Studies on Automatic Speaker Recognition

Xue YANG

B.E.
Beijing Univ. of Posts & Telecommunications
P.R.China
and
M.E.
Univ. of Electronic Science & Technology of China
Chengdu, P.R.China

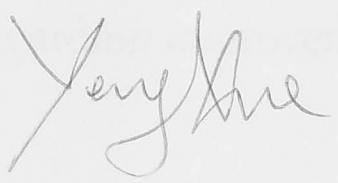
A thesis submitted for the degree of
Doctor of Philosophy
of the Australian National University

Computer Sciences Laboratory
Research School of Information Sciences and Engineering
The Australian National University

October 1997

Declaration

The contents of this thesis are the results of my own original research, and have not been submitted for a higher degree at any other university or institution.



Xue YANG

Acknowledgements

I would like to express my gratitude to the following persons and organizations.

Dr. Bruce Millar, my supervisor, for his rigorous requirements to my Ph.D. study, which has trained me in scientific research skills, and for offering criticism, suggestions and discussion time regarding improvements to draft versions of this thesis.

Prof. Richard Brent, the head of Computer Sciences Laboratory for all the support he has provided to me. Without his support, I couldn't finish this thesis. Drs. Iain Macleod, Neville Fletcher and Tony Burkitt for their concern and encouragement.

Fangxin Chen for his help in collecting speech data. Anthony Hyde, Bruce Millar, Dave Davies, Graeme Cornish, Iain Macleod, John Wach, Kevin Robert, Neville Fletcher, Steve Battisson, Tony Burkitt, Walter Bagnarol who are the speakers of my monosyllabic English word database, for generously spending the time and providing their speech data.

Frantz Clermont, Simon Hawkins, Shuping Ran, Danqing Zhang, Andrew Tridgell and Ahmady Satriawan, former fellow students at Computer Sciences Laboratory, with whom I have shared the happiness with achievements and the sadness with sufferings in our thesis work, for the understanding, moral support and technical help they gave me.

Michelle Moravec, Dennis Andriolo, Joe Elso, Arthur McGuffin for their kindness and constant technical support, and all other members of the Computer Sciences Laboratory for their friendship during my years there.

Maohui Yang, my father, and Lishan Xue, my mother, for the lasting love

and support, and Yaping He, my husband, for his love, understanding, patience and support.

The Australian National University for the PhD Scholarships.

Abstract

This thesis reports studies related to automatic speaker recognition. The studies have tackled some problems in this area from both an engineering point of view and a scientific point of view.

In the study on speaker modeling for the speaker recognition, the problems associated with two basic issues (choice of model form and modeling method) for statistical model-based speaker recognition are addressed. The mathematical relations between the VQ model and the Gaussian Mixture model are clearly revealed. An innovative modeling method has been developed to improve the ranking order of probabilities between speakers' models, and hence to improve the performance of speaker identification tasks when the model form selected is not completely correct.

In the study on the sources of inter- and intra- speaker variability in the acoustic dynamics of speech, another source of the variability at the articulatory level is investigated. The study's results lead to a conclusion that articulatory dynamic aspects (e.g. speed of tongue or lip movements), as well as articulatory static aspects (e.g. size and shape of the vocal tract), are the sources of variability in the acoustic dynamics of real speech signals.

In the study on the separation of speech signal variances from two sources, a novel method based on hidden Markov modeling techniques is developed to separately model the monotonic acoustic dynamics due to phonetic variation and the variance due to speaker variation. The method can be used in scientific analysis of dynamic speech events and their accompanying speaker variance. The associated experimental investigations suggest that the different kinds of

sources should be clarified before applying the acoustic variance measures in practical systems.

Contents

Declaration i

Acknowledgements ii

Abstract iv

1 Introduction 1

1.1 Motivation 1

1.2 Thesis Outline 3

2 Speech Materials and Transformations 7

2.1 SHLRC_SCRIBE Database 7

2.2 The Monosyllabic Word Database 8

2.2.1 The Speech Materials 8

2.2.2 Data Collection 10

2.2.3 Data Labeling 13

2.3 Data Transformation 13

3 On Speaker Modeling for the Speaker Recognition 15

3.1 Introduction 15

3.2 Comparison of VQ & GM Models 16

3.2.1 Introduction 16

3.2.2 Mathematical Analysis 17

3.2.3 Experimental Analysis 20

3.3 On the Modeling Approaches 23
3.3.1 Introduction 23
3.3.2 Review of the Discriminative Training Methods 24
3.3.3 The Speaker Identification Experiments 26
3.3.4 Alternatives to MLE Method 30
3.3.5 The Experimental Results 34
3.4 Summary 40

4 On the Sources of Inter- & Intra- Speaker Variability in the Acoustic Dynamics of Speech 42
4.1 Background 42
4.2 Aims of the Study 45
4.3 Mapping from Articulatory Parameters to Acoustic Parameters 47
4.4 Mapping from Time to Articulatory Configurations - Articulatory Dynamics 52
4.5 Boundary Conditions of Model Parameters for the Simulated Sounds 57
4.6 Verification of Simulated Sounds by Listening 60
4.7 Experimental Results: The Acoustic Dynamics 61
4.7.1 Tongue Moving from Low-back to High-front 62
4.7.2 Tongue Moving from Low-back to High-back with Lips from Unrounding to Rounding 66
4.7.3 Tongue Moving from High-back to High-front with Lips from Rounding to Unrounding 70
4.8 Summary 75

5 On the Separation of Speech Signal Variances from Two Sources 77
5.1 Introduction 77
5.2 Methods 81
5.2.1 The Model and Its Structure 82
5.2.2 The Criterion for Estimation 84

5.2.3 The Algorithm 84
5.2.4 Initial Model 85
5.2.5 Choice of the Number of States N 86
5.3 Experimental Investigations 89
5.4 Summary 94

6 Concluding Discussions **97**

6.1 On Speaker Modeling for the Speaker Recognition 97
6.1.1 Choice of Model Form 97
6.1.2 The Modeling Approach 98
6.2 On the Sources of Inter- & Intra- Speaker Variability in the Acoustic Dynamics of Speech 101
6.3 On the Separation of Speech Signal Variances from Two Sources 103
6.4 Automatic Speaker Recognition Research from Two Perspectives 105
6.5 Contributions of the Thesis 106

A Model Parameters for the Simulated Sounds **109**

B The Figures of the Global Variance and Intra-speaker Repetition Variance **119**

Bibliography 130

Chapter 1

Introduction

1.1 Motivation

Over the last two decades, automatic speaker recognition, which is defined as speaker recognition from the speech signals by machine, has become a popular topic in speech research. This topic has attracted many researchers with different backgrounds: speech science, computer science and engineering, physics and electrical engineering. Similar to automatic speech recognition, the research in automatic speaker recognition can be divided into two groups: one is the engineering research, the other is the scientific research.

The engineering research currently dominates the speaker recognition research. It is concerned with the implementation of automatic speaker recognition tasks, which are usually classified into speaker identification and speaker verification. In speaker identification, speech samples from an unknown speaker is analysed and compared with models of known speakers. The unknown speaker is identified as the speaker whose model best matches the input speech samples. In speaker verification, an identity claim is made by or asserted for the unknown speaker. The unknown speaker's speech sample is compared with the model for the speaker whose identity is claimed. If the match is good enough, as indicated by passing a threshold test, the identity claim is verified. There are many potential applications of these tasks in daily life. The

research efforts are being made to develop useful techniques and improve the task performance. The techniques cover: for example, selection of the acoustic features, the use of different model forms, the choice of speech data for modeling and the threshold for making decision *etc.* The identification performance is evaluated by the identification error rate. The verification performance is evaluated by the type I and type II error rates, where the type I error refers to false rejection of the genuine speaker's claim and the type II error refers to false acceptance of an impostor's claim. Comprehensive reviews on the techniques, tasks performance and practical considerations for automatic speaker recognition can be seen in papers by Atal (1976), Doddington (1985), O'Shaughnessy (1986) and Rosenberg *et al.* (1991), Furui (1994).

In the scientific research, some fundamental problems regarding the relationships between speaker's characteristics and his speech are enquired. For example, there is a qualitative discussion on what the sources of inter- and intra-speaker variability are and how they are manifested in the acoustic signals (Stevens, 1971). The attempts were also made to qualitatively model the transformations occurring at different levels of human speech production: semantic, linguistic, articulatory and acoustic (Nolan, 1983).

Regarding the relations between the scientific and engineering research in automatic speaker recognition, it is natural to think that the scientific research discovers insight and provides solid knowledge about how humans produce and perceive speech signals, and that the engineering research uses this knowledge and develops methods to construct the practical system. However in the real world of speech research area, we are facing two facts. On the one hand, our knowledge about human speech communications are, as described by Fant (1989), "*in many respects incomplete and primitive*". "*We may have acquired a reasonable overall qualitative view of the nature of speech, but we have not been able to organize our insights into quantitative, operative representations*" (Fant, 1989) – the quantitative representations required in implementing a practical system. It is most likely due to this fact that, some reefs on the

way towards the destination island of the speaker recognition tasks still can not be got over. On the other hand, given limited knowledge about human speech communications, the engineering research seeks for the solutions from mathematics. The mathematical statistic methods and their innovations are introduced to quantitatively and effectively describe the general nature of the speech signals. A good example is the use of hidden Markov model to represent the dynamic and static natures of the speech signals. Without doubt, these techniques are also useful tools for the scientific research of automatic speaker recognition.

Based on a general review of the automatic speaker recognition literature, my conclusion is that there are still many problems in both understanding the human speech communication process and developing a reliable recognition system. The two groups both need to continue, and they can inspire each other on the way to the destination island of automatic speaker recognition tasks we expect. Working in this multidisciplinary area, speech researchers should have the scientific and engineering knowledge on the relevant topic. With all these in mind, during my Ph.D. study, I have made attempts to attack the problems of automatic speaker recognition from scientific and engineering points of view, and meanwhile to get myself trained in these two aspects. This thesis reports the details of my work.

1.2 Thesis Outline

There are another five chapters and two appendixes following this chapter. They together make up this thesis.

Chapter 2 describes the speech materials used in this thesis work and the acoustic pre-processing procedure for these materials.

I first picked up the engineering problems without hesitation, because they are more popular than the scientific problems. The statistical model based speaker recognition approach is the most accepted one in the engineering re-

search (this approach is simply called model based approach hereafter in the thesis). The basic idea of the approach is that a known speaker is represented by the probability distribution of his speech signals. In the speaker identification, the recognition rule is the well-known Bayesian decision procedure. That is, the recognition decision is made according to the maximum value of the following joint probabilities (where $i = 1, ..., M$) among M known speakers included in the system,

$$maxP(S_i, O) = maxP(O|S_i)P(S_i)$$

In this equation, $P(S_i)$ is the probability of speaker S_i, $P(O|S_i)$ is the probability of the speech signal O from the speaker S_i. If we assume the speaker probability distribution over the given speaker set is uniform, then only the probability distribution of the speech signals from each known speaker is needed. In the speaker verification system, it requires $P(O|S_i)$ to compare with a threshold before the decision is made. Therefore the probability distribution of the speech signals from a known speaker is essential for such model based speaker recognition systems.

It should be noted that, under the model-based approach paradigm, there are two implied assumptions: 1) individual speaker has his own probability distribution of the speech signals; the distributions of different speakers are different to each other, 2) for each speaker, the training data which are used in establishing the model and the test data which are used in recognition phase, follow the same probability distribution. In reality, these two assumptions may not be true, and because of it, recognition errors can be introduced.

To implement a speaker recognition system, the real distribution has to be estimated from the speech signals and represented by a mathematical function. If an estimated distribution (a model) can correctly represent the real distribution, the system is optimal in the sense of minimum error rate. Chapter 3 reports the study on speaker modeling for the speaker recognition, associated with choice of model form and methods of estimation of model's parameters

from the speech signals of a known speaker for better speaker recognition performance.

In scientific approach to automatic speaker recognition, some basic questions are asked. What are the sources of inter- and intra-speaker variability in the speech signal? How do they individually influence the speech signals? These are difficult questions, but the answers to them would help to build up a speaker recognition system in which each model is able to capture more features in the speech of a speaker and able to adapt to the speech variance caused by apparent physiological and/or emotional change of the speaker. Currently, the discussions related to these questions are at primitive stage, basically following the categorization first made by Garvin & Ladefoged (1963) that speaker variability in the speech signals arises from organic (anatomical, physiological, structural) and learned (acquired, behavioral, functional) differences.

Chapter 4 reports the study on the sources of inter- and intra- speaker variability in the acoustic dynamics of speech. Appendix A includes some of the materials associated with this study. In the study, I looked at the sources at the articulatory level and their effects on the dynamic aspects of the first three resonant frequencies of some dynamic speech sounds.

Variability is an inherent characteristic of speech signals. The acoustic variance within one utterance over time can be the realisation of intentional phonetic variation. The variance over repeated utterances of the same speech events can be caused by inter- and/or intra-speaker variation. Additional sources of acoustic variance external to the speaker may also be present such as that due to the recording environment, but these we may regard as secondary.

Chapter 5 reports the study on the separation of speech signal variances from two sources, and Appendix B contains part of the results. The sources considered are intra-speaker variation and phonetic variation. To identify the sources of acoustic variance and to separately describe the variances from different sources are important contributions to both scientific and engineering research into automatic speaker recognition. The contribution to scientific

research is concerned with the acoustic nature of different variances. The contribution to engineering research is concerned with the design of ways to apply that nature to the speaker recognition task.

Broad discussions and conclusions on the results of the studies described in Chapter 3, 4, and 5 are given in the last chapter, Chapter 6.

Chapter 2

Speech Materials and Transformations

Two databases were involved in this thesis work. Some speech materials from SHLRC_SCRIBE database were used in the study on speaker modeling for the speaker recognition. The Monosyllabic Word database was designed and used mainly in the study on the separation of speech signal variances from two sources: phonetic variation and speaker variation, and associated with the study on the sources of inter- and intra- speaker variability in the acoustic dynamics of speech. In the following sections, these databases and their associations with the studies are described.

2.1 SHLRC_SCRIBE Database

SHLRC_SCRIBE database is part of Australian National Database of Spoken Language (ANDOSL) project. ANDOSL project is a cooperative database development project sponsored by the National Spoken Language Database committee of the Australian Speech Science and Technology Association, and implemented in its initial stages by four institutions (Sydney University, the National Acoustic Laboratories, Macquarie University, and the Australian National University) with funding support from Australian Research Coun-

cil (Millar *et al.*, 1994). SHLRC_SCRIBE database uses the speech materials taken from set A, B, C, D and P of SCRIBE (Hieronymus *et al.*, 1990) project. The speech data were collected and labeled at the Speech Hearing and Language Research Centre (SHLRC), Macquarie University, Australia (Croot *et al.*, 1992). Set A consists of 200 sentences, each produced by five male, general Australian speakers. Among them, the 22 sentences labeled as from 030 to 051 by each of the speakers were used in the study of this thesis on speaker modeling for the speaker recognition. The context of each of these sentences is listed in Table 2.1. Generally, these 22 sentences cover the whole phonemic inventory of Australian English.

2.2 The Monosyllabic Word Database

2.2.1 The Speech Materials

The speech materials of this database were designed to include some monosyllabic English words which show their acoustic dynamics. I started with the articulatory domain to choose the word list. The tongue and lips in the vocal tract are only considered, as their movement is responsible for the production of most speech sounds. The word list was chosen such that to produce each word, the tongue traverses a certain space from one end to the other of the vocal tract and/or the lips move from "rounded" to "unrounded" or vice versa. Consequently, the corresponding acoustic representations of each word become dynamically variable. Table 2.2 lists the words (i.e. 7 common English monosyllabic words) selected for the database.

This data corpus was designed to be spoken 32 times by 11 Australian male speakers in 4 recording sessions. The target time interval between sessions was designed to be 1 week with each word being spoken 8 times in each session by each speaker. It was intended that each speaker pronounce these words over the sessions in normal physiological and emotional states. Hence, only

030	opportunities like this don't grow on trees as you should know by now.
031	it's difficult to choose between two such equally good alternatives.
032	henry normally viewed rob's zealous enthusiasm with disdain.
033	she'll think of an excuse if given enough space and time.
034	when forced to make a choice sarah chose ping-pong as her favourite game.
035	it's a shame that architects design for themselves and not for the general public.
036	coe beat him to the line by five-thousandths of a second.
037	the bath plug is missing so you'll have to take a shower.
038	you ought to brush your teeth before you go to bed.
039	the earth used to be flat but now it's a sphere.
040	I wish he'd either grow a beard or shave his moustache.
041	judith found the manuscripts waiting for her on the piano.
042	the huge castle was encircled by a deep moat.
043	jane adored maths and french but hated the rest of school.
044	mashed potatoes are more fattening than either boiled or baked ones.
045	the world is becoming increasingly dangerous but hardly anyone cares.
046	patrick will need speech therapy because of his cleft palate.
047	these practical jokes have been taken much too far.
048	I shall paint this room mauve with a few beige dots.
049	my car was stolen while I was shopping on kensington high street.
050	changing gear half way up a steep hill can be quite risky.
051	I always enjoy a pint of lager when I come of the squash court.

Table 2.1: The sentences in SHLRC_SCRIBE database used in the thesis work.

word
we
you
how
high
sauce
judge
pap

Table 2.2: The word list of the Monosyllabic Word database.

one form of intra-speaker variance (that associated with repetition of the same word on different occasions), which is referred to as intra-speaker repetition variance hereafter, is another main feature of this database.

2.2.2 Data Collection

The data were collected in an acoustically treated speech laboratory of Computer Sciences Laboratory, Research School of Information Sciences and Engineering, Australian National University. During recording, the speaker seated in front of the Intel 486 personal computer. A headset microphone (Model PPS-2A) was put about 1 cm from the corner of his mouth rather than at centre of the mouth. The microphone is connected to a low-pass filter (Rockland Model 432) with a bandwidth as 9 kHz. The signal from the filter is connected to a sound card installed in the computer.

The sound card was set up via a control program, part of which was developed by the TRUST (Technology for Robust User-conscious Secure Transactions) project affiliated with Computer Science Laboratory, Research School of Information Science and Engineering, Australian National University. Its sampling rate was set as 20,000 samples/second and digitization was 16 bits/sample.

Prior to recording of each session, a test run of actual recording was simulated for each word by each speaker. The purpose of a test run was to determine an adequate recording level for the volume setting and to get speaker familiar with his task. The volume setting for the filter was checked/adjusted as such that the maximum intensity of the filter output for the normal-voice does not exceed $0db$ too much on the VU meter. The volume setting for the sound card was checked/adjusted via the control program to make sure the normal-voice-recorded waveform prior to A/D conversion is not clipped or energy level is not too small. In order to meet the design for the features of the database, apart from the same physical recording environment over sessions, in each session every speaker was checked to be in normal health condition (e.g. no cold caught, no dryness of the mouth and throat and no dramatic physical exercise before coming to recording) and in neutral emotion (e.g. no grief, no anger, no fear, no sorrow, no happiness). Every speaker was instructed to speak naturally.

During the recording procedure, the control program prompted the speaker to pronounce a word randomly selected from the word list and stored the collected data for each utterance into a file during each recording session.

In the course of the data collection, several speakers selected were not able to attend on schedule. As a result, minimum and maximum time intervals between successive sessions were respectively 1 day and 12 days, with an average time interval between successive sessions of 6.5 days. Occasional speaking errors made by two speakers meant that they had less than 32 correct utterances for some words — speaker IM has 31 for *high*, and speaker JW has 31 for *we*, and 30 for both *you* and *high*. A summary of the collected data is given in Table 2.3. These deviations from the original design are relatively minor and should not affect the overall results of the studies.

The collected data were down-sampled to 10,000 samples/second before further processing. This is because the most information of these data concerned in this thesis work is below 5kHz and the down-sampled data take less memory of storage.

speaker	repetitions						
	we	*you*	*how*	*high*	*sauce*	*judge*	*pap*
AH	32	32	32	32	32	32	32
BM	32	32	32	32	32	32	32
DD	32	32	32	32	32	32	32
GC	32	32	32	32	32	32	32
IM	32	32	32	31	32	32	32
JW	31	30	32	30	32	32	32
KR	32	32	32	32	32	32	32
NF	32	32	32	32	32	32	32
SB	32	32	32	32	32	32	32
TB	32	32	32	32	32	32	32
WB	32	32	32	32	32	32	32

Table 2.3: A summary of data collection results.

2.2.3 Data Labeling

The vocalic part of each utterance was manually labeled and only this part was used for the thesis work. The labeling was conducted based on visual observation of the speech waveforms and spectrograms and auditory verification in the XWAVES environment of Entropic Signal Processing System (ESPS). The labeling criteria were,

onset of the vocalic portion: starting point of periodic waveform of the utterance, which corresponds to beginning of clear formant structure.

offset of the vocalic portion: end point of periodic waveforms of the utterance, which corresponds to end of clear formant structure.

2.3 Data Transformation

A short period of the speech signals (15 - 20 ms) can be approximately regarded as having stationary spectral property. If the signals of that period are assumed to be produced from an all-pole linear time invariant system, a set of polynomial coefficients of the transfer function can be estimated from the signals by the linear prediction analysis. Those coefficients are also called linear prediction coefficients (LPCs). They effectively represent the power spectrum of the signals in that period. Further, a set of the cepstral coefficients derived from LPCs represents the corresponding log spectrum. Comprehensive discussions on this speech signal analysis technique can be found in quite a few books (e.g. Markel & Gray 1976, Rabiner 1978).

The cepstral coefficients have been widely adopted in speech/speaker recognition systems to represent the speech signals. In the thesis studies with SHLRC_SCRIBE database and the Monosyllabic Word database, the cepstral coefficients (which are also called LPC_cepstrals or cepstrals in this thesis) were used. Hence, in this common parametric space, some problems in the speech/speaker recognition research can be discussed.

The cepstrals were derived by using the programs in ESPS. Table 2.4 sum-

marises the parameters used in the transformation from the speech signals to the cepstrals. They were applied to the speech materials in both databases for the thesis work.

For SHLRC_SCRIBE database, total number of LPC_cepstrals sets for each speaker are listed in Table 2.5, reflecting the length of all sentences together and ranging from 60.15 to 72.89 seconds.

pre-emphasis:	0.98
order of LPC cepstrals:	12
window length:	20ms
window type:	Hamming
window step:	10ms

Table 2.4: A summary of transformation parameters.

speaker	number of the cepstrals
DB	6015
DW	7289
JC	6510
MB	6745
ND	7070

Table 2.5: Total number of the cepstrals sets of the materials of SHLRC_SCRIBE database for each speaker used in the thesis work.

Chapter 3

On Speaker Modeling for the Speaker Recognition

3.1 Introduction

As we have known, in model-based speaker recognition, a model to represent the real distribution of the speech signals from a speaker is required. In practice, there are two basic issues in obtaining the model for a speaker. One is the choice of model form (mathematical functional form) and the other is the criterion and its corresponding method, to estimate the parameters of the chosen model form from training samples (both together are often called the modeling approach, or estimation approach, or training approach). Which model form can better fit to the real distribution of the speech signals from a speaker? Which modeling method can lead to better recognition performance? Why is model form A (modeling method A) better than model form B (modeling method B)? Many attempts made in speaker modeling to improve the speaker recognition performance can be regarded as relating to one of these two issues. This chapter first reports the study associated with the choice of model form (Section 3.2). Then, it reports the study associated with the estimation method (Section 3.3). Finally it gives a summary about these studies.

3.2 Comparison of VQ & GM Models

3.2.1 Introduction

Vector quantisation (VQ) is a popular model form chosen for the speaker recognition (Soong *et al.* 1985, Rosenberg *et al.* 1987). It has been interpreted as a model which "*provides an effective means for compressing the short term spectral representation of speech signals*" and "*explicitly attempts to model representative short-term spectral characteristics*" (Rosenberg *et al.* 1987). Recently, Gaussian Mixture (GM) model has also been applied to speaker recognition (Reynolds 1992, Reynolds *et al.* 1995). "*The use of Gaussian mixture models for modeling speaker identity is motivated by the interpretation that the Gaussian components represent some general speaker-dependent spectral shapes and the capacity of Gaussian mixtures to model arbitrary densities*" (Reynolds *et al.* 1995). The comparison between the VQ and the GM is usually experimentally demonstrated in terms of the speaker recognition performance, where the number of codewords in the VQ is set equal to the number of components in the GM (Matsui & Furui 1994, Reynolds *et al.* 1995). The results show that GM model is superior to VQ model. Reynolds *et al.* (1995) interpret the differences between the VQ and the GM as that the VQ "*models the distributions by representative templates from hard partitions of the feature space*" and the GM "*generalizes on this notion by providing a soft partitioning of a speaker's space using Gaussian basis functions*".

In order to further understand why the GM model outperforms VQ model in the speaker recognition task, in this section, the relationship between the VQ and GM is first mathematically explicated. Based on their mathematical relations, it is then experimentally demonstrated how the VQ is a less-fit to the real speech data of a speaker than the GM.

3.2.2 Mathematical Analysis

Vector quantisation technique was initially developed for signal coding (Linde *et al.* 1980) and only later applied to speaker modeling for speaker recognition (Soong *et al.* 1985, Rosenberg *et al.* 1987). In a VQ model, a codebook comprising N codewords is formed for a speaker. Each codeword takes the form of a vector defined in the chosen parametric space. In the training phase (speaker modeling), after an initial specification of codewords is made, the "K-means" algorithm of cluster analysis of statistics (e.g. Hartigan, 1975) is used to iteratively estimate the value of each codeword. With the algorithm, the data samples are assigned to the clusters according to their nearest codewords, and each codeword is refined as the mean of samples allocated to the cluster. This procedure continues until minimisation of the cumulative encoding error for the given training samples is achieved.

The Gaussian mixture model is most commonly reported in the literature on statistical cluster analysis. When modeling a speaker, it is assumed that data $\mathbf{x}_t$ $(t = 1, ., T)$ from the speaker follows a composite distribution which can be described by a weighted sum (or mixture) of N Gaussian distributions,

$$p(\mathbf{x}_t|\lambda) = \sum_{n=1}^{N} c_n \mathcal{N}_n(\mathbf{x}_t; \mu_n, \mathbf{\Sigma}_n) \tag{3.1}$$

where $\mathcal{N}_n(\mathbf{x}_t; \mu_n, \mathbf{\Sigma}_n)$ is a multivariate Gaussian probability density function with the form,

$$\mathcal{N}_n(\mathbf{x}_t; \mu_n, \mathbf{\Sigma}_n) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}_n|^{1/2}} exp(-\frac{1}{2}(\mathbf{x}_t - \mu_n)^{T_r} \mathbf{\Sigma}_n^{-1}(\mathbf{x}_t - \mu_n)) \tag{3.2}$$

Parameters $\lambda = \{c_n, \mu_n, \mathbf{\Sigma}_n, n = 1, ..., N\}$ summarise a Gaussian mixture model (d in Equation 3.2 is the dimension of $\mathbf{x}_t$). They are respectively the mixture weight, the mean vector, and the covariance matrix for component n.

Expectation/Maximisation (EM) algorithms (e.g. Everitt & Hand, 1981) are mostly used to iteratively estimate the parameters λ in maximum likelihood sense. Given an initial model λ^i, the E (Expectation) step first computes the posterior probability that the observation $\mathbf{x}_t$ belongs to component n,

$$P(n|\mathbf{x}_t, \lambda^i) = \frac{c_n \mathcal{N}_n(\mathbf{x}_t; \mu_n, \mathbf{\Sigma}_n)}{\sum_{n_1=1}^{N} c_{n_1} \mathcal{N}_{n_1}(\mathbf{x}_t; \mu_{n_1}, \mathbf{\Sigma}_{n_1})}; n = 1, ..., N. \tag{3.3}$$

Then, M (Maximisation) step re-estimates the parameters $\lambda^{(i+1)} = \{c_n^{(i+1)}, \mu_n^{(i+1)}, \Sigma_n^{(i+1)}, n = 1, ..., N\}$. For each component n,

$$c_n^{(i+1)} = \frac{1}{T} \sum_{t=1}^{T} P(n|\mathbf{x}_t, \lambda^i) \tag{3.4}$$

$$\mu_n^{(i+1)} = \frac{\sum_{t=1}^{T} P(n|\mathbf{x}_t, \lambda^i) \mathbf{x}_t}{\sum_{t=1}^{T} P(n|\mathbf{x}_t, \lambda^i)} \tag{3.5}$$

$$\Sigma_n^{(i+1)} = \frac{\sum_{t=1}^{T} P(n|\mathbf{x}_t, \lambda^i)(\mathbf{x}_t - \mu_n)(\mathbf{x}_t - \mu_n)^{T_r}}{\sum_{t=1}^{T} P(n|\mathbf{x}_t, \lambda^i)} \tag{3.6}$$

It is shown below that VQ model is simply a constrained or degenerate case of more complex GM model.

1. **Non-overlapping constraint**. If the components in a GM model are well separated, the probability density value of $\mathbf{x}_t$ on an exclusive component will be much larger than the values on other components. In this case, the posterior probability that $\mathbf{x}_t$ is from that exclusive component is near unity, while this probability with respect to other components is near zero,

$$P(n = n_e|\mathbf{x}_t, \lambda^i) = \frac{c_{n_e} \mathcal{N}_n(\mathbf{x}_t; \mu_n, \mathbf{\Sigma}_n)}{\sum_{n=1}^{N} c_n \mathcal{N}_n(\mathbf{x}_t; \mu_n, \mathbf{\Sigma}_n)} \doteq \frac{c_{n_e} \mathcal{N}_n(\mathbf{x}_t; \mu_n, \mathbf{\Sigma}_n)}{c_{n_e} \mathcal{N}_n(\mathbf{x}_t; \mu_n, \mathbf{\Sigma}_n)} = 1 \tag{3.7}$$

$$P(n \neq n_e|\mathbf{x}_t, \lambda^i) = \frac{c_n \mathcal{N}_n(\mathbf{x}_t; \mu_n, \mathbf{\Sigma}_n)}{\sum_{n=1}^{N} c_n \mathcal{N}_n(\mathbf{x}_t; \mu_n, \mathbf{\Sigma}_n)} \doteq \frac{c_n \mathcal{N}_n(\mathbf{x}_t; \mu_n, \mathbf{\Sigma}_n)}{c_{n_e} \mathcal{N}_n(\mathbf{x}_t; \mu_n, \mathbf{\Sigma}_n)} = 0 \tag{3.8}$$

This constraint is directly comparable with VQ modeling, where K-means algorithm partitions the data samples into separate clusters. Each $\mathbf{x}_t$ is allocated to an exclusive cluster with probability unity, namely, the cluster to which $\mathbf{x}_t$ has minimum distance. In this respect, the VQ resembles the GM under the constraint that components are non-overlapping.

2. **Component weight constraint**. Given that there are L_n data samples allocated to component n, the non-overlapping component constraint leads to the re-estimation of c_n being

$$c_n^{(i+1)} = \frac{1}{T}\sum_{t=1}^{T} P(n|\mathbf{x}_t, \lambda^i) = \frac{L_n}{T} \tag{3.9}$$

where c_n is proportional to the number of observations allocated to component n. Putting a further constraint on the GM by fixing c_n equal to $1/N$, causes the GM to resemble the VQ in that the information on the number of data samples allocated to each cluster is disregarded.

3. **Covariance constraint**. The constraint on c_n means that the allocation of $\mathbf{x}_t$ to component n in the GM is determined only by $\mathcal{N}_n(\mathbf{x}_t; \mu_n, \mathbf{\Sigma}_n)$. If one further constraint is put on the GM, that the covariance matrix for each component n is the identity matrix (i.e. $\Sigma_n = I$), then $\mathcal{N}_n(\mathbf{x}_t; \mu_n, \mathbf{\Sigma}_n)$ becomes,

$$\mathcal{N}_n(\mathbf{x}_t; \mu_n, I) = \frac{1}{(2\pi)^{d/2}} exp(-\frac{1}{2}(\mathbf{x}_t - \mu_n)^{T_r}(\mathbf{x}_t - \mu_n)) \tag{3.10}$$

This constrains that $\mathcal{N}_n(\mathbf{x}_t; \mu_n, I)$ is the function only of the Euclidean distance between $\mathbf{x}_t$ and μ_n. Since the exponential function in $\mathcal{N}_n(\mathbf{x}_t; \mu_n, I)$ is inversely proportional to the Euclidean distance, allocation of the data sample $\mathbf{x}_t$ therefore can be based on minimum distances rather than probability densities. This is exactly how the VQ partitions the data samples.

When the above three constraints are applied to GM model, the only parameters left to be estimated are the mean vectors $\lambda = \{\mu_n, n = 1, ..., N\}$.

This estimation takes the following form:

$$\mu_n^{(i+1)} = \frac{\sum_{t=1}^{T} P(n|\mathbf{x}_t, \lambda^i)\mathbf{x}_t}{\sum_{t=1}^{T} P(n|\mathbf{x}_t, \lambda^i)} = \frac{\sum_{\mathbf{x}_t \in n} \mathbf{x}_t}{L_n}; n = 1, ..., N \tag{3.11}$$

where L_n is the number of observations allocated to component n. This is the same method as used in the VQ for generation of codewords.

3.2.3 Experimental Analysis

The above mathematical analysis has shown that GM model with the addition of three constraints: non-overlapping between components, equal mixture weights and identity of covariance matrix of each component, degenerates into VQ model. Are these three constraints applicable to real-world speech signals? The experimental analysis described below provides some evidences to answer this question.

The basic idea of the experimental analysis is that it is assumed the three constraints are applicable to real speech data. That is, the three constraints are exerted on GM model (equivalent to VQ model) to represent the data. Then it is to examine whether the estimated GM models follow the assumption. If they did, the assumption would be valid; otherwise it would be invalid.

The speech materials of the SHLRC_SCRIBE database by each of the five speakers as described in Chapter 2 were used in the analysis. The number of codewords (called components hereafter) N in a model was set to 32, one of the codebook sizes chosen in most speaker recognition research.

First, it was examined whether the components in a model are overlapping or not. The *posterior probability*, that each sample of the same speech data as used in modeling belongs to component n, was calculated using Equation 3.3 in the previous sub-section, where each c_n was set to $\frac{1}{32}$ and each Σ_n was set to I. Table 3.1 lists the number of samples for each of which the maximum value of $P(n|\mathbf{x}_t, \lambda), n = 1, ..., N$ falls in six different ranges.

speaker	N_0	N_m					
		[.0,.1)	*[.1,.2)*	*[.2,.3)*	*[.3,.4)*	*[.4,.5)*	*[.5,1.]*
DB	6015	5012	888	115	0	0	0
DW	7289	5540	1251	379	117	2	0
JC	6510	5130	1062	283	35	0	0
MB	6745	5265	1277	203	0	0	0
ND	7070	5937	1019	114	0	0	0

Table 3.1: N_m, number of data samples for each of which the maximum values of $P(n|\mathbf{x}_t, \lambda)$ with the GM model falls in six different ranges. N_0, total number of data samples tested.

It can be seen from Table 3.1 that, for each speaker, no data sample has a *posterior probability* greater than 0.5 of falling in one of 32 components. Most data samples belong to one particular component with maximum probability less than 0.1. It can be concluded that components of the GM models in the experiments heavily overlap.

Second, it examined the values of mixture weight $c_n, n = 1, ..., N$, using Equation 3.9 of the previous sub-section. The minimum of c_n, c_{min} and the maximum of c_n, c_{max} are listed in Table 3.2 for each speaker. It can be seen that c_{max} is at least five times greater than c_{min}.

Third, It was examined whether the covariance matrix of each component is effectively an identity matrix.

Each of 32 covariance matrices is computed based on the data samples allocated to the corresponding component. Each of 12 elements in diagonal of 32 covariance matrices is examined, and the largest one among them for each speaker is listed in Table 3.3. Although it was not examined whether each correlation value off the diagonal of 32 covariance matrices is near to zero or not, the values in Table 3.3 are much less than one, indicating that identity covariance matrix is not valid for each component for all speakers.

speaker	c_{min}	c_{max}
DB	5.486284e-03	5.087282e-02
DW	1.783509e-03	5.364247e-02
JC	5.376344e-03	5.330261e-02
MB	1.482580e-04	6.137880e-02
ND	1.074965e-02	5.417256e-02

Table 3.2: Maximum and minimum values of the mixture weight c_n in the GM model of each speaker.

speaker	σ^2_{max}
DB	0.102259
DW	0.092330
JC	0.091259
MB	0.094840
ND	0.093924

Table 3.3: The largest value σ^2_{max} of diagonal elements in 32 covariance matrices in the GM model of each speaker.

The experimental results demonstrate that none of the three constraints on the GM is applicable to the real speech of each of the five speakers tested, suggesting that GM model is more suitable than VQ model to represent the speech signals of a speaker.

3.3 On the Modeling Approaches

3.3.1 Introduction

Given a model form selected for a speaker, the maximum likelihood criterion with its corresponding method, Maximum Likelihood Estimation (MLE), is often used to estimate the model parameters (e.g. Reynolds *et al.* 1995). The method is an iterative procedure to obtain the model parameters λ_i, for speaker i by making the following probability function move towards its local maximum value (X_i is the observation sequence from speaker i used as training data and M is total number of models in the system).

$$P(X_i|\lambda_i) \qquad i = 1, ..., M$$

If the model form is correct, the real distribution can be represented as close as possible by the model estimated by the MLE and the identification error rate can be eventually minimised. In an attempt to improve the recognition performance, correctness of the selected model forms has been challenged (e.g. Bahl *et al.* 1988, 1990, Ephraim & Rabiner 1990). How can we modify the model estimation method to move the speaker identification performance towards minimum error rate, if the selected model form is not completely correct? In this section, some issues relating to this question are discussed.

Sub-section 3.3.2 first reviews two well-known so-called discriminative training approaches developed for the speech or speaker recognition: the maximum mutual information estimation (MMIE) method and the corrective training method. These two approaches were proposed to replace MLE method in case an incorrect model form was used. The review ends up with three

questions: 1) how does the incorrect model form affect the recognition performance? 2) do these two methods necessarily lead to the models which can improve performance of the system with the models estimated by the MLE? 3) If 2) is true, how do they perform that improvement?

Speaker identification experiments based on GM model form and MLE method have been conducted, to investigate how the incorrect model form could affect the recognition performance. Sub-section 3.3.3 reports the experimental results, provides an answer to the question 1) and raises the proposal and hypothesis based on the experimental results.

The MMIE and the corrective training methods have been analysed in an attempt to answer the question 2) & 3). Based on the proposal, an innovative modeling approach has been developed to reduce the recognition errors, which are assumed to be due to incorrectness of the model form. The details of these are in Sub-section 3.3.4.

The experimental results using the developed method and their relations to the hypothesis and questions raised, are provided in Sub-section 3.3.5.

3.3.2 Review of the Discriminative Training Methods

In the speech recognition research, the model form is assumed (e.g. hidden Markov model) and MLE method is used to find the parameters of the model that best fit to the real distributions(e.g. Rabiner 1989). It has been argued, however, that our current understanding of speech is at a primitive stage and correctness of model form selected for any word is not always true (Bahl *et al.* 1988, 1990, Ephraim & Rabiner 1990). In this situation, a speech recognition system built on MLE modeling approach is sub-optimal.

In attempts to optimise the performance of the speech recognition system, MMIE approach has been proposed by the speech recognition group at IBM (Bahl *et al.* 1986, 1990) to estimate the models with the criterion to discriminate between the likelihood the word is produced from its own model and the likelihood it is produced from other words' model. Similar ideas have also been

proposed by the researchers at AT & T Bell Lab (Ephraim, Dembo & Rabiner 1989, Ephraim & Rabiner 1990, Ljolje, Ephraim, Rabiner 1990). Formulated in hidden Markov models, MMIE approach estimates parameters of models λ_i, $\lambda_j, j = 1..M, j \neq i$ to maximise each statistic,

$$S_i = \frac{p(\mathbf{X}_i|\lambda_i)}{\sum_{j=1,j\neq i}^{M} p(\mathbf{X}_i|\lambda_j)} \qquad i = 1, ..., M \tag{3.12}$$

$\mathbf{X}_i$ is the observation sequence from word i rather than from word j. The MMIE tries to maximise $p(\mathbf{X}_i|\lambda_i)$ (the same as MLE approach), and at same time to minimise $\sum_{j=1,j\neq i}^{M} p(\mathbf{X}_i|\lambda_j)$. The estimation is usually performed using a general optimisation procedure, which is also iterative to locally maximise $S_i, i = 1...M$. The reported experimental results show that the performance of a speech recognition system trained by MMIE approach is better than that trained by MLE approach (Bahl *et al.* 1986, Bahl *et al.* 1990). The mathematical comparison between the MMIE and the MLE (Nadas, *et al.* 1988) also claimed that, in minimising the error rate of Bayes-decision based speech recognition, MMIE approach "*may be preferable when the model is incorrect*".

After investigating MMIE approach, the speech recognition group at IBM proposed another training method, corrective training method (Bahl *et al.* 1988), which was inspired by the error-corrective training procedure for linear classifiers. Developed in discrete hidden Markov model based speech recognition, the corrective training further adjusts probability values of a model initially estimated by the MLE, by 1) using mis-recognised utterances in the training data and applying one iteration of estimation of MLE method to make those utterances more probable from the model, and 2) subtracting a certain quantity from the model of the other word as of which the utterances were mis-recognised to make those utterances less probable from that model. Minimising error rate of the training data is used as criterion in model estimation. Their experimental results show that, the corrective training leads to better recognition performance than the MMIE. But "*given the poorly understood nature of corrective training, it is difficult to say why it was found to be so*

much more effective than MMIE' (Bahl *et al.* 1988).

Similar to the situation in speech recognition research, our knowledge about underlying probabilistic structure of the data from any speaker is limited. With the motivation to improve the recognition performance, researchers at AT& T Bell Lab. have tried a discriminative training method in hidden Markov model based speaker recognition (Liu, Lee, Juang & Rosenberg 1994). The basic idea of the method is similar to the MMIE as it also intends to discriminate between the likelihood that the signal is generated from speaker's own model and the likelihood it is generated from other speakers' models.

The methods described above were developed for the case where the model form is not assumed to be completely correct and use of MLE method can only reach a sub-optimal performance. Three questions arise after this review: 1) how does the incorrect model form affect the recognition performance? 2) can these two methods necessarily improve performance of the system with models estimated by the MLE? 3) If 2) is true, how do they perform that improvement?

3.3.3 The Speaker Identification Experiments

The aim of this sub-section is to describe the speaker identification experiments and their results. The experiments have been conducted to investigate how incorrect representation of real distributions could affect the recognition performance.

The experiments used the speech materials of the SHLRC_SCRIBE database, described in Chapter 2 from the given five speakers set, and GM models with the number of components N for each model being 32. The parameters of GM models were estimated by MLE approach. In the experiments, the data samples used for modeling were also used for testing in order to guarantee that training data and test data follow the same distribution and that the errors caused by different distributions between the training data and test data are avoided. Therefore, there were only two types of recognition errors involved, if

any; i.e. the errors due to overlapping between the distributions and the errors due to incorrect model form used. Recognition rate was recorded for each set of cepstrals (each sample) of the data. The ovearall recognition results for each speaker are shown in Table 3.4.

speaker	N_0	N_c	$N_e(\%)$
DB	6015	4728	1287 (21.40)
DW	7289	5750	1539 (21.11)
JC	6510	5506	1004 (15.42)
MB	6745	5199	1546 (22.92)
ND	7070	5381	1689 (23.89)

Table 3.4: The experimental results based on the GM models estimated by the MLE. N_0, total number of data samples tested. N_c, the number of samples which are correctly recognised. N_e, the number of samples which are mis-recognised. The value in bracket (the error rate) after N_e is the percentage of N_e in N_0.

The tested data samples were then classified into two groups: Group c and Group e. Group c consists of the samples (N_c) which were correctly recognised, and Group e consists of the samples (N_e) which were mis-recognised. Further for each speaker i, the mean M of the minimum difference of the log probability density between his own model and other speakers' models over N_c samples in Group c, was calculated as follows,

$$M = \frac{1}{N_c} \sum_{n=1}^{N_c} min\{p[\mathbf{x}_i(n)|\lambda^i] - p[\mathbf{x}_i(n)|\lambda^j], j = 1, ..., 5; j \neq i\} \qquad (3.13)$$

where $p[\mathbf{x}(n)_i|\lambda^i]$ is the log probability density value of a data sample $\mathbf{x}(n)$ from speaker i on his own model λ^i and $p[\mathbf{x}(n)_i|\lambda^j]$ is the log probability density value of a data sample $\mathbf{x}(n)$ from speaker i on the model λ^j of speaker j. The results are in Table 3.5.

speaker	N_c	M
DB	4728	4.7
DW	5750	5.1
JC	5506	6.3
MB	5199	4.2
ND	5381	3.9

Table 3.5: N_c, total number of data samples which are correctly recognised. M, the mean over N_c, of the minimum difference of log probability density between the speaker's own model and other speakers' models.

For each data sample in Group c of each speaker, the minimum difference, in log probability density between the speaker's own model and the other speakers' models was calculated,

$$d_m = min\{p[\mathbf{x}_i(n)|\lambda_i] - p[\mathbf{x}_i(n)|\lambda_j], j = 1, ..., 5; j \neq i\} \qquad (3.14)$$

It can be seen from Table 3.6 that there is approximately 60% of the data samples ($N_c^{'}$) in Group c of each speaker, for which their d_m goes beyond quantity δ (= 2.5 which is less than M).

For each data sample in Group e of each speaker, the absolute difference, in log probability density between the speaker's own model and the mis-recognised foreign model was calculated,

$$d_e = |p[\mathbf{x}_i(n)|\lambda^i] - p[\mathbf{x}_i(n)|\lambda^j]| \qquad (3.15)$$

Table 3.7 shows that there is approximately only 20% of the data samples ($N_e^{'}$) in Group e of each speaker, for which their d_e goes beyond quantity δ (= 2.5 which is less than M).

With $\delta = 2.5(< M)$, the experimental results clearly indicate the difference between $N_c^{'}$ and $N_e^{'}$ in percentage of the corresponding group.

speaker	N_c	$N_c^{'}$ (%)
DB	4728	3239 (68.5)
DW	5750	3795 (66.0)
JC	5506	4207 (76.4)
MB	5199	3237 (62.3)
ND	5381	3112 (57.8)

Table 3.6: N_c, the number of samples of Group c. $N_c^{'}$, the number of samples in Group c for which the minimum difference in log probability density between the speaker's own model and the other speakers' models is equal to or larger than δ. The value in bracket after $N_c^{'}$ is the percentage of $N_c^{'}$ in N_c. ($\delta = 2.5$)

speaker	N_e	$N_e^{'}$ (%)
DB	1287	263 (20.4)
DW	1539	350 (22.7)
JC	1004	210 (20.9)
MB	1546	326 (21.1)
ND	1689	343 (20.3)

Table 3.7: N_e, the number of samples in Group e. $N_e^{'}$, the number of samples which are mis-recognised and for which the difference in log probability density between other speakers' model as mis- recognised and the speaker's own model is equal or larger than δ. The value in bracket after $N_e^{'}$ is the percentage of $N_e^{'}$ in N_e. ($\delta = 2.5$)

It should also be noted that the Bayes rules used in the recognition phase means that the ranking order of probabilities between speakers rather than the absolute probability values essentially determines the final decision. In the experiments, the estimated GM model may follow general contour of the real distribution, but may not fit to its details. Therefore if the mismatch between the real distribution of a speaker's data and the estimated distribution with GM model form causes recognition errors, those errors occur more likely in the regions, in which the differences between a speaker's own distribution and other speakers' distributions are relatively small and ranking relations of these distributions are not correctly represented by the estimated models.

As a consequence of the experiments, it is proposed that we may be able to reduce those errors by improving the ranking order between the estimated models but sacrificing fitness of those models to the real distributions. We may increase the data samples of each speaker in the regions, in which the difference in probability between his own model and one of the other speakers' models are relatively small, to enhance detail modeling of the distribution in an attempt to improve correctness of the ranking order between the models in those regions. To increase the data samples in these regions will decrease the probabilities in other regions because sum of probabilities over the whole region should be one according to the mathematical definition. If this decrease does not affect the ranking relations between models, we may be able to achieve the purpose. Further a hypothesis is proposed that the models better fit to the real distributions in the sense of maximum likelihood such as those estimated by MLE method may not necessarily lead to higher performance if the model form is not completely correct.

3.3.4 Alternatives to MLE Method

In this sub-section, the MMIE and the corrective training method are first analysed in terms of their relations to improvement of correctness of ranking order between models of speakers, hence to improve the recognition performance.

Then it reports an innovative modeling approach which has been developed based on the proposal raised in the previous sub-section.

In the MMIE, we need to estimate parameters $\lambda_i, \lambda_j; j = 1, ..., M,$ $j \neq i$, to maximise each statistic $S_i, i = 1, ..., M$ as in Equation (3.12). For each S_i, because $p(X_i|\lambda_i), p(X_i|\lambda_j)$ $(i \neq j)$ are independent functions, we maximise S_i by maximising $p(\mathbf{X}_i|\lambda_i)$ and $\frac{1}{p(\mathbf{X}_i|\lambda_j)}(j = 1, ..., M, j \neq i)$ individually. Eventually, we need to estimate $\lambda_i, i = 1, ..., M$ by maximising each of following item simultaneously,

$$p(\mathbf{X}_i|\lambda_i); \qquad \frac{1}{p(\mathbf{X}_j|\lambda_i)} \qquad j = 1, ..., M, j \neq i$$

If we start with $\lambda_i, i = 1, ..., M$ estimated by the MLE, where $p(\mathbf{X}_i|\lambda_i)$ has a single maximum (this is more likely a case with the real speech data), and then use the MMIE to re- estimate, the final estimated result of λ_i will be the same as that estimated by the MLE. The MMIE could make the statistics S_i greater than that obtained in the MLE only when $p(\mathbf{X}_i|\lambda_i)$ has more than one maximum. Even so, it can not be seen how the MMIE explicitly improves the ranking order in some particular regions (i.e. the regions in which the ranking order of probability between speakers is not correctly represented). Does the greater statistic S_i (which is a measure of discrimination degree between models of speakers) defined in the MMIE (see Equation 3.12) produce the models which necessarily lead to less errors in the recognition?

The corrective training method adjusts probabilities of models in all regions in which any mis-recognised training data fall, in order to make those data more probable from their own model (by increasing the value in the probability function) and less probable from the foreign model as of which they are mis-recognised (by decreasing the value in the probability function). This method may perform improvement of the ranking order of the probabilities in part of those regions between models. But it should be noted that not all errors are caused by incorrect model form (there are inherent errors due to overlapping of real distributions between speakers). The method was developed intuitively

following the error correcting training procedure for linear classifiers. Because of this, no explanations on why the particular techniques were chosen and how the method worked effectively as reported, have been provided by the developers.

However, the corrective training method suggests us that improvement of the correct ranking order can be directly examined by the recognition rate of the training data. Based on the proposal in the previous sub-section, an innovative modeling approach has been developed as follows.

In developing the innovative modeling method, two specific questions were to be answered. 1) what regions do we select for each model to add more data samples in training? 2) how do we add more data samples to enhance modeling of the detail in order to improve the ranking order of probabilities between models in the selected regions?

The regions are selected in which the training data fall and their absolute difference between log probability densities produced by its own model and by the closest model of other speakers is less than a certain quantity δ. The choice of δ is dependent on data distributions of the given speaker set. If δ is too small, some confusable regions may be ignored, in which recognition errors occur due to inaccurate modeling. If δ is too large, more regions will be selected, which have the correct ranking order between models. The more these regions are selected, the less effective it is to improve the ranking order in the confusable regions. Eventually, the choice of δ will be judged by recognition error rate of the training data. The method does not follow the corrective training in that all regions, in which the training data are mis-recognised, are selected, because some errors are caused by the inherent ranking order relations of probabilities between real distributions.

Unlike the corrective training method which applies one iteration of the MLE to the selected training data to re-adjust the parameters of the model, the method adds the selected training data samples to the training set used for the previous model estimation to form a new training set for each speaker.

This new training set is then used to update the parameters of the model by MLE approach. The new training set includes the data used for the previous training because the updated model should still reflect the feature of data distribution in the other regions. Comparing to the previous model, the updated model will have greater probability in the selected regions and less probability in the other regions. The updated model is poorer fit to the real distribution because it is estimated by using exaggerated training set. However, here it is trying to improve correctness of the ranking order of probabilities between models in the confusable regions by sacrificing correct fitness to real distribution. The recognition error rate of the genuine training data is used to judge this improvement. It is different from the corrective training in which a certain quantity of the probability is also subtracted from other speakers' models which are close to the model in the selected regions. This subtraction is actually equivalent to addition of the probability to speaker's own model. It works only in a way to possibly speed up adjustment towards correct ranking order of probabilities between models.

The answers to the two questions can be summarised as two steps which form the innovative modeling method for speaker identification.

Step 1: for speaker i, given model λ_i^k estimated from the data samples set $\sum_i^k$, the following is calculated for each data sample $\mathbf{x}_i(n), n = 1, ...N_0$ which is produced by speaker i,

$$min\{|p(\mathbf{x}_i(n)|\lambda_i) - p(\mathbf{x}_i(n)|\lambda_j)|, j = 1, ..., 5; j \neq i\} \tag{3.16}$$

the data samples whose value of (3.16) is less than δ are collected to form a separate data sample set $\sum_i$.

step 2: $\sum_i^k$ and $\sum_i$ are merged to form a new training data set $\sum_i^{k+1}$ which will be used to update model λ_i^k to λ_i^{k+1} by the MLE.

These two steps can be iteratively performed starting with initial models derived by the MLE using $\mathbf{x}_i(n), n = 1, ...N_0$. The recognition rate among samples $\mathbf{x}_i(n), n = 1, ...N_0$ produced by speaker i can be used as the objective

function to control the number of iterations for this method.

3.3.5 The Experimental Results

The method has been applied to model speech materials of five male speakers in SHLRC_SCRIBE database with GM model form in the same way as that in the experiments described in Sub-section 3.3.3. Again, the data samples used for modeling were also used for testing to guarantee that training data and test data follow the same distribution and that no errors due to different distributions between the training data and test data are introduced. We believe that if we had separate training and test data sets for each speaker and their distributions were identical, the recognition results would be the same as what we have obtained with single data set. In each iteration, the recognition test is performed by using the genuine training set (the same as the training set used in the MLE to form the initial models). Table 3.9, Table 3.10, Table 3.11 and Table 3.12 show the recognition results after each of four iterations with $\delta = 2.5$. For comparison purpose, the recognition results on the initial models as shown in Table 3.4 are listed again in Table 3.8.

From these five tables, it can be seen that, for all speakers but ND, the error rate decreases with increase of number of iterations. Although speaker ND has the error rate fluctuating with the number of iterations, the results still show that the method has improved correctness of the ranking order of probability between the models estimated by the MLE, thus has improved the identification performance. Table 3.13 lists, for each speaker, the number of samples by which the recognition errors are reduced with the models trained by 3 iterations of the innovative modeling method. In the bracket after each value in the table, it is the percentage of error reduction in the total errors of Table 3.8. The recognition errors in Table 3.8 include two types of errors: one caused by inherent overlapping of distributions; the other caused by inaccurate representation of probability ranking among the distributions. It can be seen from Table 3.13 that, for example, speaker DB has at least 184 errors (14.30%

speaker	N_0	N_c	N_e^0 (%)
DB	6015	4728	1287 (21.40)
DW	7289	5750	1539 (21.11)
JC	6510	5506	1004 (15.42)
MB	6745	5199	1546 (22.92)
ND	7070	5381	1689 (23.89)

Table 3.8: Recognition results with the GM model estimated by the MLE. N_0, total number of data samples tested. N_c, the number of samples which are correctly recognised. N_e^0, the number of samples which are mis-recognised. The value in bracket (the error rate) after N_e^0 is the percentage of N_e^0 in N_0.

speaker	N_0	N_c	N_e^1 (%)
DB	6015	4854	1161 (19.30)
DW	7289	5879	1410 (19.34)
JC	6510	5651	859 (13.19)
MB	6745	5288	1457 (21.60)
ND	7070	5452	1618 (22.88)

Table 3.9: Recognition results after iteration 1 of the innovative estimation of the GM model. N_0, total number of data samples tested. N_c, the number of samples which are correctly recognised. N_e^1, the number of samples which are mis-recognised. The value in bracket (the error rate) after N_e^1 is the percentage of N_e^1 in N_0.

speaker	N_0	N_c	N_e^2 (%)
DB	6015	4903	1112 (18.49)
DW	7289	5889	1400 (19.21)
JC	6510	5728	782 (12.01)
MB	6745	5320	1425 (21.13)
ND	7070	5447	1623 (22.96)

Table 3.10: Recognition results after iteration 2 of the innovative estimation of the GM model. N_0, total number of data samples tested. N_c, the number of samples which are correctly recognised. N_e^2, the number of samples which are mis-recognised. The value in bracket (the error rate) after N_e^2 is the percentage of N_e^2 in N_0.

speaker	N_0	N_c	N_e^3 (%)
DB	6015	4912	1103 (18.34)
DW	7289	5916	1373 (18.84)
JC	6510	5782	728 (11.18)
MB	6745	5348	1397 (20.71)
ND	7070	5473	1597 (22.59)

Table 3.11: Recognition results after iteration 3 of the innovative estimation of the GM model. N_0, total number of data samples tested. N_c, the number of samples which are correctly recognised. N_e^3, the number of samples which are mis-recognised. The value in bracket (the error rate) after N_e^3 is the percentage of N_e^3 in N_0.

speaker	N_0	N_c	N_e^4 (%)
DB	6015	4951	1064 (17.69)
DW	7289	5939	1350 (18.52)
JC	6510	5824	686 (10.54)
MB	6745	5359	1386 (20.55)
ND	7070	5458	1612 (22.80)

Table 3.12: Recognition results after iteration 4 of the innovative estimation of the GM model. N_0, total number of data samples tested. N_c, the number of samples which are correctly recognised. N_e^4, the number of samples which are mis-recognised. The value in bracket (the error rate) after N_e^4 is the percentage of N_e^4 in N_0.

errors among the total errors in Table 3.8), which belong to the second type of the error and which have been eliminated by using the innovative modeling method with 3 iterations.

The facts, that the error rate for speaker ND after iteration 2 (see Table 3.10) is higher than that after iteration 1 and the error rate after iteration 4 (see Table 3.12) is higher than that after iteration 3, could be explained as follows. For those samples that changed to be mis-recognised after iteration 2 and 4 of using the method, the re-estimated models produce the sample with probability from the greatest to non-greatest, which directly affects the performance. If the error rate is regarded as the objective function of the method, it can be seen that its value does not always decrease monotonically with number of iterations as the objective function in MLE method does. More iterations may lead to a form of over-trained models which can not represent correct ranking orders of probability in their real data distributions.

In order to test the hypothesis proposed at the end of Sub-section 3.3.3 and to answer the question raised when analysing the MMIE method in Sub-section 3.3.4, $\ln p(\mathbf{X}_i|\lambda_i)$ and $\ln p(\mathbf{X}_i|\lambda_j), i, j = 1...5, j \neq i$, have been com-

speaker	N_e^0 - N_e^3 (%)
DB	184 (14.30)
DW	166 (10.79)
JC	276 (27.49)
MB	149 (9.64)
ND	92 (5.45)

Table 3.13: The number of samples, by which the recognition errors are reduced with the models trained by 3 iterations of the innovative modeling method. The value in bracket is the percentage of error reduction in the total errors of Table 3.8.

puted for the model of each speaker i estimated by the MLE and the model estimated by the innovative modeling method after 3 iterations. They are listed in Table 3.14 and Table 3.15 respectively. It can be seen that each diagonal element in Table 3.14: $p(\mathbf{X}_i|\lambda_i)$, is greater than the corresponding value in Table 3.15. In almost all cases (apart from DB's data on JC's model), $\sum_{j=1,j\neq i}^{M} p(\mathbf{X}_i|\lambda_j)$ in Table 3.14 is less than its counterpart in Table 3.15. These indicate that, 1) the models estimated by the innovative modeling method are poorer fit to the real distributions in maximum likelihood sense than the models estimated by the MLE, 2) the discrimination in the sense of maximum mutual information as defined in the MMIE between the models estimated by the innovative modeling method is poorer than that between the models estimated by the MLE. However, as shown in Table 3.8 and Table 3.11, the recognition error rate based on the models by the innovative modeling method is lower than that based on the models by the MLE.

speaker	Probability density values on a model				
	$\ln p(\cdot\|\lambda_1)$	$\ln p(\cdot\|\lambda_2)$	$\ln p(\cdot\|\lambda_3)$	$\ln p(\cdot\|\lambda_4)$	$\ln p(\cdot\|\lambda_5)$
DB$\rightarrow \mathbf{X_1}$	33261.47	-2245.98	-18638.5	-6981.78	-5664.17
DW$\rightarrow \mathbf{X_2}$	-19961.83	28249.31	-38278.72	-17157.26	- 22067.85
JC$\rightarrow \mathbf{X_3}$	-29961.99	-19879.61	32854.96	-30611.25	- 27242.63
MB$\rightarrow \mathbf{X_4}$	-8019.31	-4766.62	-18389.80	33683.61	287.13
ND$\rightarrow \mathbf{X_5}$	-5918.19	-10493.47	-15598.40	-3359.14	32543.28

Table 3.14: Probability density values of $\mathbf{X}_i$ for each speaker on models $\lambda_j, j = 1...5$, which are estimated by the MLE for speaker DB, DW, JC, MB, ND respectively.

speaker	Probability density values on a model				
	$\ln p(\cdot\|\lambda_1)$	$\ln p(\cdot\|\lambda_2)$	$\ln p(\cdot\|\lambda_3)$	$\ln p(\cdot\|\lambda_4)$	$\ln p(\cdot\|\lambda_5)$
DB$\rightarrow \mathbf{X_1}$	32091.90	-1481.68	-19036.89	-5834.82	- 4080.94
DW$\rightarrow \mathbf{X_2}$	-18587.09	26931.88	-36246.77	-15391.81	- 18490.10
JC$\rightarrow \mathbf{X_3}$	-27785.27	-19695.12	31377.59	-27876.87	- 25349.51
MB$\rightarrow \mathbf{X_4}$	-8954.99	-4035.44	-16422.93	32468.17	2313.57
ND$\rightarrow \mathbf{X_5}$	-5428.16	-9064.79	-15108.00	-2243.98	31614.66

Table 3.15: Probability density values of $\mathbf{X}_i$ for each speaker on models $\lambda_j, j = 1...5$, which are estimated by 3 iterations of the innovative estimation method for speaker DB, DW, JC, MB, ND respectively.

3.4 Summary

Some problems relating to the two basic issues in speaker modeling for the speaker identification have been addressed: 1) choice of model form, 2) modeling method.

In the first part of this chapter, VQ and GM model forms have been compared. Rather than following the usual paradigm of assessing their speaker recognition performance, the comparison has: (i) explicated mathematical relationships between the two models, showing that VQ model is equivalent to GM model with three imposed constraints: 1) non-overlapping between components, 2) equal mixture weights, 3) identity of covariance matrix of each component, and (ii) experimentally demonstrated that the three constraints on the GM are not applicable to real speech data, suggesting GM model can provide more effective means than the VQ to model a speaker for the speaker recognition.

In the second part of this chapter, some problems associated with modeling method have been discussed.

The most popular, so-called discriminative modeling methods have been reviewed: maximum mutual information estimation method and corrective training method, which were previously developed to replace maximum likelihood estimation method and to improve the speech/speaker recognition performance when the model form used is not exactly correct for real distributions. The outcome of the review is three questions: 1) how does the incorrect model form affect the recognition performance? 2) can these two methods necessarily improve performance of the system with models estimated by the MLE? 3) If 2) is true, how do they perform that improvement?

Speaker identification experiments have been conducted in order to investigate how the incorrect model form could affect recognition performance. The suggestion from the experiments is that, if the mismatch between the real distribution of a speaker's data and estimated model causes recognition er-

rors, those errors may occur in the regions, in which the differences between a speaker's own distribution and other speakers' distributions are relatively small and ranking relations of these distributions are not correctly represented. Based on the experimental analysis, a proposal has been raised to improve correctness of the ranking order of probabilities between models.

The relations of the MMIE and the corrective training method to improvement of correctness of the ranking order between models have been analysed. It has been argued that the model by the MMIE will remain the same if it is initially estimated by the MLE and has single maximum value. The MMIE seems not to be related to that improvement, while the corrective training has its own flaws because it intuitively follows the error correcting training procedure for linear classifiers.

Based on the proposal, an innovative modeling method has been developed to improve correctness of the ranking order of probabilities between the models. The experimental results with a five speakers set and GM model form have shown more correct ranking order between the models estimated by the method than those estimated by the MLE, hence better recognition performance than those by the MLE.

It has also been demonstrated that the models by the innovative modeling method are poorer fit to the real distributions in the sense of maximum likelihood and are less discriminative in the sense of maximum mutual information than the models estimated by the MLE. Conclusions can be drawn from the results that if the model form used is not completely correct, 1) the models better fit to the real distributions such as those estimated by the MLE do not necessarily lead to less error rate, 2) the models with greater discrimination such as those estimated by the MMIE do not necessarily lead to less error rate either.

Further discussions and conclusions are given in Chapter 6.

Chapter 4

On the Sources of Inter- & Intra- Speaker Variability in the Acoustic Dynamics of Speech

4.1 Background

Automatic speaker recognition has motivated scientific enquiry on the sources of inter-speaker and intra-speaker variability in the speech signals. Perhaps, the most comprehensive but qualitative discussion on the sources of inter- and intra-speaker variability in the speech signal, is the one given by Stevens (1971). In his paper, he describes various physiological sources of inter- and intra-speaker variability in the acoustic properties of many English static sounds. In the later part of his paper, he writes (p. 223) "*Most of the acoustic characteristics that have been considered here are measured during steady-state regions of the sound output, when the vocal tract and larynx structures are in more-or-less steady configurations. It is to be expected, however, that differences or variability in articulatory dynamics account for a number of acoustic attributes that distinguish one speaker from another or that are likely to show intra-speaker variability. Inter-speaker variability in such attributes may be*

the result of learned articulatory habits rather than differences in anatomical or physiological characteristics of the articulatory structure, but nevertheless may often provide important cues for the identity of a talker". The acoustic attributes of articulatory dynamics to which Stevens referred are patterns of formant movements, particularly during liquids, glides and diphthongs, such as slopes of formants and durations.

Wolf (1972) and Sambur (1975) have reported their studies on the selection of acoustic parameters for the speaker recognition, and have discussed the relations between the sources in speech production system and the some acoustic parameters. Wolf (1972, p. 2045) writes, "*Organic differences are the result of variations in the sizes and shapes of the components of the vocal tract: larynx, pharynx, tongue, teeth, and the oral and nasal cavities. Since the resonances of the vocal tract and the characteristics of the sound energy sources depend upon just these anatomical factors, organic differences lead to differences in fundamental frequency, laryngeal source spectrum, and formant frequencies and bandwidths. Learned differences are the result of differences in the patterns of coordinated neural commands to the separate articulators learned by each individual. Such differences give rise to variations in the dynamics of the vocal tract such as the rate of formant transitions and coarticulation effects*". Sambur describes his work (1975, p. 176) as "*The investigation was conducted by first determining an initial set of acoustic parameters which might be suitable candidates for indicating the unique properties of a speaker's vocal apparatus, as well as some aspects of his learned pattern of speaking*". His experimental results demonstrated that the important features in recognising an unknown speaker include the formant structure of vowels which "*is directly related to the unique shape of the vocal tract and supplies important information about the speaker's identity*" and the measurements such as slope of the second formant in /aI/, which are "*related to the dynamic properties of the talker's voice patterns that reflect his learned behavior of speaking*".

In review papers on the speaker recognition, most authors inevitably men-

tion the sources of inter- and intra-speaker variability in the speech signal.

Atal writes (1976, p. 461) "*Speaker-related variations in speech are caused in part by the anatomical differences in the vocal tract and in part by the differences in the speaking habits of different individuals. The anatomical differences relate to the fixed structural differences in the shape or size of the vocal tract which can vary considerably from one person to another. The differences in the speaking habits, on the other hand, result from the manner in which persons have learned to use their speech mechanism. Such differences show up in the temporal variation of speech characteristics of different individuals.*"

When talking about acoustic features, O'Shaughnessy (1986, p.12) writes "*the two sources of speaker variation, physiological and behavioral differences, lead to two types of useful features. Inherent features are relatively fixed for a speaker and depend upon the anatomy of his vocal tract. ... , inherent features are less susceptible to the mimicry of impostors than learned features. The latter refers to the dynamics of vocal tract movement – i.e., the way a speaker talks.*"

Rosenberg *et al.* (1991 b, p.702) write "*Broadly speaking, speaker identity is correlated with physiological and behavioral characteristics of the speaker*". Following that, they give several examples of acoustic outputs due to physiological differences, one of which is that "*variations in the size of vocal tract cavities produce differences in the characteristic resonances of the spectrum of speech signals*". They then continue, "*Low-level behavioral information is associated with individual characteristics in articulating speech sounds, in characteristic transitions from sound to sound, in characteristic pitch contours, rhythm, and timing, and so forth*".

The statements quoted above represent current understanding of the origins of speaker variability in the speech signals and lead us to the knowledge that 1) there are two basic sources at the articulatory level that reflect the speaker's organic and learned differences: articulatory structure and articulatory dynamics. Articulatory structure is dictated by the shape and size

of speaker's vocal apparatus, while articulatory dynamics is dictated by the movement and coordination of the articulators in the vocal tract, 2) the acoustic static speaker variability directly relates to the variability in articulatory structure, while the acoustic dynamic speaker variability directly relates to the variability in articulatory dynamics.

Although no systematic investigations have been designated to answer questions about the sources of speaker variability in the speech signal, the previous studies focusing on English vowel sounds provide some evidences which can form a partial basis for our knowledge of the sources of speaker variability. Peterson and Barney (1952) have reported the first two resonant frequencies (formants) distributions for nine vowels spoken by men, women and children. Stevens & House (1963) and Stevens (1971) have measured the vocal tract length for each of three speakers, averaged over four conditions of articulations, and the first three formants for the each speaker, averaged over his eight vowels. Their results lead to a conclusion that as the vocal tract cavities get larger, the formant frequencies decrease. Based on their 3-parameter lossless acoustic tube model of the vocal tract, Stevens & House (1955) have given detailed results suggesting the relations between a variety of vocal tract shapes and their formants for English vowels. The results can help to understand how different vocal tract shapes generate the formants for a static vowel sound. Regarding the source of variability in speech dynamics, it seems natural to immediately point to articulatory dynamics. However, a careful consideration about the process of speech production may give rise to a question: is articulatory dynamics the only such source, or does articulatory structure also contribute to this variability?

4.2 Aims of the Study

The speech signal is the output of a speech production system. Any acoustic parameter of the speech signal as a function of time is actually formed by two

mappings: one mapping from time to the configuration of the system and one mapping from the configuration to the acoustic parameter.

The mapping from time to the configuration of the system relates to articulatory dynamics. It is the movement of articulators in the system, such as tongue body and lips, that make changes in the configuration, which in turn makes changes in the speech signal. There is little quantitative evidence about articulatory dynamics, i.e. how configuration of the system changes with time during the articulatory realisation of the speech. But it can be clearly assumed that articulatory dynamics may vary from speaker to speaker and from time to time. The variations will result in different dynamics in the speech signal. Thus, the articulatory dynamics is one possible source of inter- and intra-speaker variability in the acoustic dynamics of speech.

The mapping from the articulatory configuration to the acoustic parameter is determined by articulatory structure. Using a model-based approach, Stevens &House (1955) and Fant (1960) have reported the detailed relations between constriction position and the formants under different conditions of constriction degree and lip rounding. When constriction degree and lip rounding are fixed, the effects of constriction position on formants (displayed as formant trajectories which are called nomograms by Fant), show a nonlinear property (see Fant 1960 pp. 82-84). This property has been later interpreted in terms of the quantal nature of speech and in turn related to distinctive features in speech sound systems (Stevens 1972, 1989). This nonlinearity basically represents how much and in what manner formants change as the result of tongue body movement along the vocal tract length specified by the change of constriction position. These characteristics are necessarily included in the corresponding formant trajectories as a function of time as tongue body moves with time. The effects of constriction degree and lip rounding on the formant trajectories as a function of constriction position do not simply move the trajectories upwards or downwards, they show nonlinearity too (see Fant 1960 pp. 82-84).

Based on these observations, we propose a hypothesis that differences in size and shape of the vocal tract also influence the dynamics of the formant trajectories in the speech signal.

This chapter reports the study to test this hypothesis. Three dynamic phonetic segments have been simulated by performing the two mappings. They are /aI/ as in "high", /aU/ as in "how" and /wi/ as in "we". In producing these three sounds, tongue body movement traverses the vocal tract cavity and lip rounding also varies. The implementation of the mapping from the parameters of the vocal tract model to the first three resonant frequencies is addressed in next section. The mapping from time to the parameters of the model for each sound is discussed in Section 4.4. Controlled perturbation to the model's structure have been applied to represent different size and shape of the vocal tract for each sound. Section 4.5 reports how to choose the values of the model's static parameters and boundary values of time-varying parameters for each sound and how to vary some static parameters' values and time-varying parameters' boundary conditions. In order to confirm that the simulated sounds retain their intended phonetic quality, the synthesised waveforms of each simulation of the sounds was listened by the author to perform an informal verification (see Section 4.6). Section 4.7 analyses in detail the relations of different sizes and shapes of the vocal tract model to the dynamic aspects of the formant trajectories of these sounds. A summary is given in Section 4.8.

4.3 Mapping from Articulatory Parameters to Acoustic Parameters

The vocal tract has its unique geometric structure. It has been approximated by acoustic lossless tube models (Fant 1960, Stevens & House 1955) to study the relations between the vocal tract shape and the acoustic outputs for vowels

and consonants. Fant (1960) published his book representing the comprehensive studies on the topic, and the results have been widely accepted and quoted by speech research community. Mermelstein (1973), Coker (1976) and Wakita & Fant (1978) demonstrated more complicated models to produce more natural speech sounds. The sounds synthesized in our study fall in the vowel space in which the static vowels have been thoroughly examined by Fant using his acoustic lossless tube model. Therefore, the model chosen for the study is based on Fant's (1960) model. One configuration example of the model is shown in Figure 4-1. Eight parameters employed to define a model configuration are,

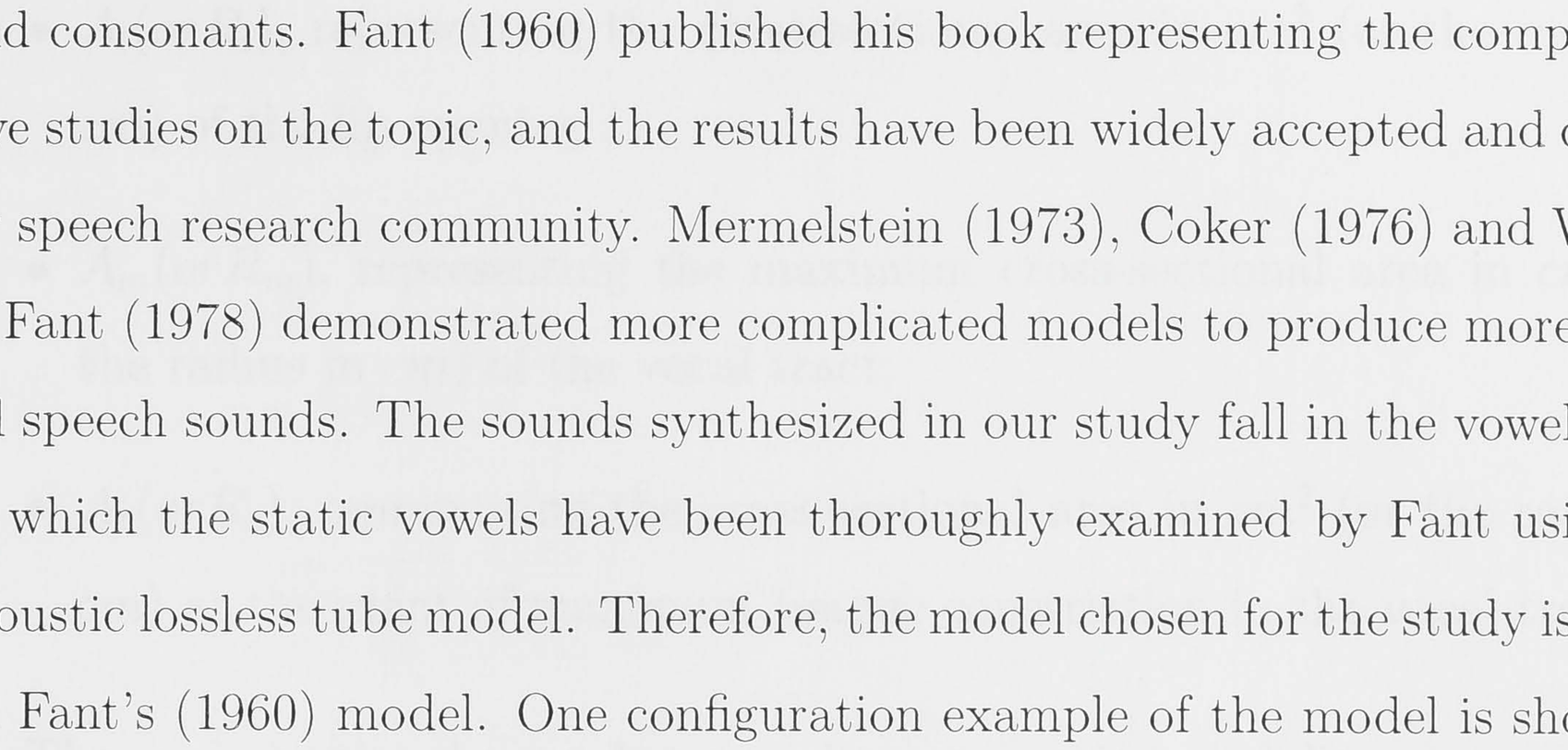


Figure 4-1: One example of the model with a particular configuration, which shows cross-sectional area $A(x)$ in cm^2 against distance from the glottis x in cm.

- L_g, representing the length in cm of the larynx portion;
- L_m, representing the length in cm of the lip section;
- L, representing the length in cm of the vocal tract;
- X_c, representing the distance in cm of the maximum tongue constriction position to the glottis;
- A_g, representing the cross-sectional area in cm^2 of the larynx portion;

- $A_l(or R_l)$, representing the cross-sectional area in cm^2 (or the radius in cm) of the lip opening ;

- $A_m(or R_m)$, representing the maximum cross-sectional area in cm^2 (or the radius in cm) of the vocal tract;

- $A_c(or R_c)$, representing the cross-sectional area in cm^2 (or the radius in cm) at the point of maximum tongue constriction in the vocal tract.

The cross-sectional area between larynx portion and lip section in the model, $A(x)$ as a function of distance from the glottis x is defined by

$$A(x) = A_c \times cosh^2(\frac{x - X_c}{h})$$

where

$$h = \frac{4.75cm}{arccosh\sqrt{\frac{A_m}{A_c}}}$$

As a constant value, h is determined by the condition that the area shall always reach A_m at $\pm 4.75cm$ from X_c and then remain constant at this value in both the front and the back cavity.

The area functions for larynx portion and lip section are constant.

The above acoustic lossless tube model can be further quantised as a concatenation of a finite number of short uniform sections of equal length as in Figure 4-2. The cross-sectional area for each section is the average of the function over the short length of that section.

One configuration of the quantised version of the model specified by those articulatory parameters defined above, determines how air flow transmits in the model. Spectral properties of the configuration at the lips end can be mathematically described by its transfer function. The transfer function, the poles of which are the resonant frequencies, has a form of rational function in the Z domain of digital signal processing with the denominator being a polynomial in Z^{-1}: $D_N(Z) = \sum_{i=0}^{N} a_i Z^{-i}$ (this type of transfer function is also called all-pole transfer function). It is actually a function of geometric

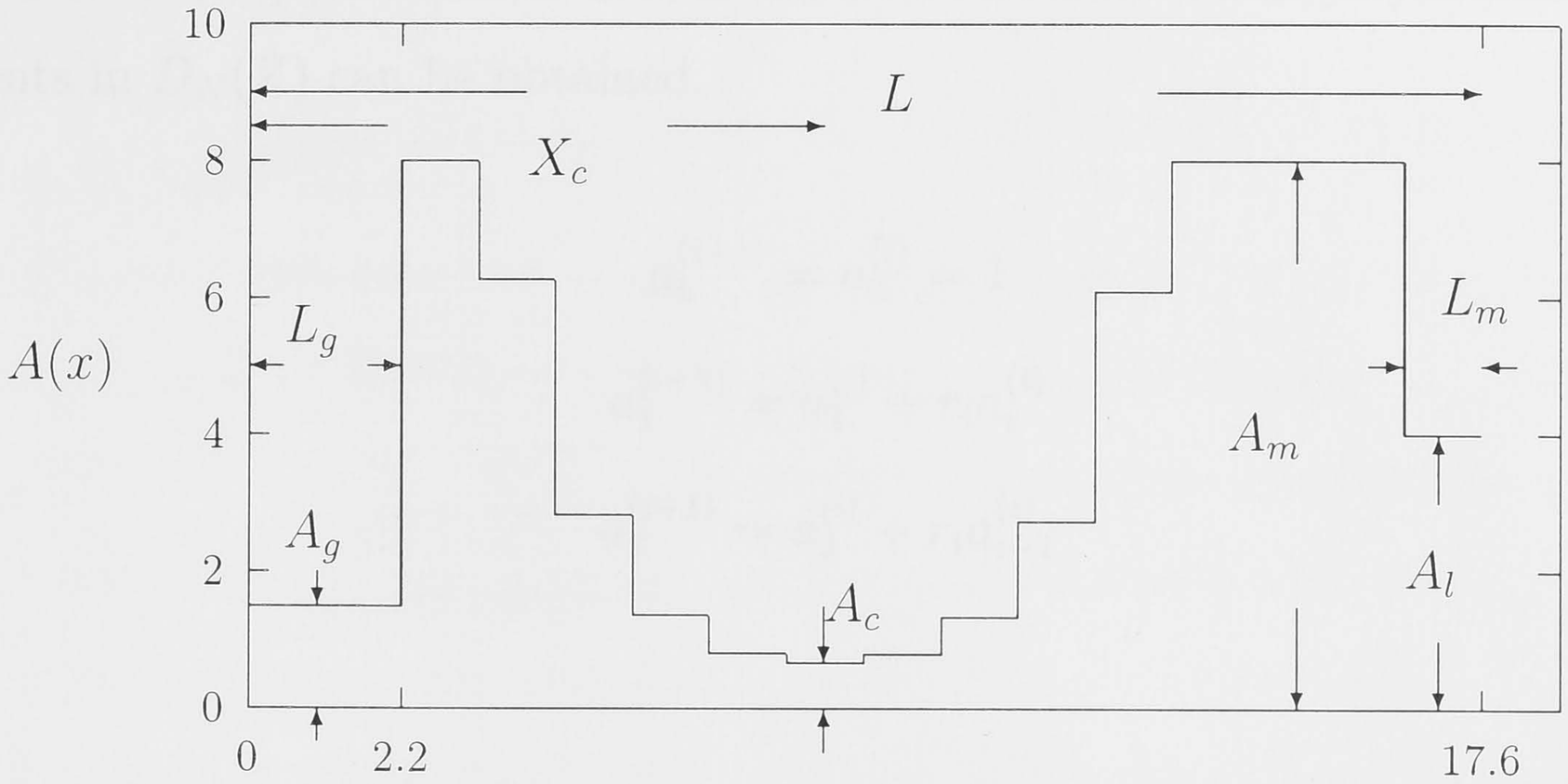


Figure 4-2: One example of digitised model configurarion, which shows cross-sectional area $A(x)$ in cm^2 against distance from the glottis x in cm. Each short section is labeled as $i, i = 1, ..., N$, starting from the glottis.

parameters of the model (so are the resonant frequencies). The relation of a set of cross-sectional areas of the model to a set of polynomial coefficients and the derivation of the poles from the polynomial coefficients have been discussed in quite a few references (e.g. Markel & Gray 1976, Rabiner 1978).

Here the procedure, used in the experiments, of derivation of the resonant frequencies from a model configuration defined by a set of cross-sectional areas, is summarised as follows,

- The area for the section $i, i = 1, ..., N$ is represented by A_i, see Figure 4-2.
- $r_i = \frac{A_{i+1}-A_i}{A_{i+1}+A_i}$ represents the percentage of travelling waves in the tube that is reflected at the junction between i and $i+1$ and called the reflection coefficient at that junction.
- The wave travelling in the lossless tube of Figure 4-2 can be mathematically described by a concatenation of a number of lattice filters. Thus, $D_N(Z)$ can be derived recursively,

$$D_0(Z) = 1$$

$$D_k(Z) = D_{k-1}(Z) + r_k Z^{-k} D_{k-1}(Z) \qquad k = 1, 2, ..., N$$

Further, the relations of the reflection coefficients to the polynomial coefficients in $D_N(Z)$ can be obtained,

$$a_0^{(i+1)} = a_0^{(i)} = 1$$

$$a_1^{(i+1)} = a_1^{(i)} + r_i a_i^{(i)}$$

$$a_2^{(i+1)} = a_2^{(i)} + r_i a_{i-1}^{(i)}$$

$$\cdots$$

$$a_i^{(i+1)} = a_i^{(i)} + r_i a_1^{(i)}$$

$$a_{i+1}^{(i+1)} = r_i$$

• In digital signal processing, when the signal is assumed to be produced by the system with an all-pole transfer function (the same as that for the acoustic lossless tube model of the vocal tract), the corresponding power spectrum can be derived from the estimated a_i. The spectrum is called the maximum entropy estimated spectrum. The algorithm used for this purpose can also be employed to derive the spectral properties of the lossless acoustic tube model used in this study. Because there exist the explicit mathematical relations between the reflection coefficients and the polynomial coefficients, the ESPS provides a program to derive the power spectrum directly from the reflection coefficients. This program was used in the study.

• The model used in this study is limited in frequency by the length of the uniform section. Its bandwidth is determined by $f_0 = \frac{c}{4\times\Delta l}$, where c is the velocity of sound in $cm/sec.$, Δl is the length of each section of the model in cm. In the following experiments, Δl is chosen as 1.1 and the bandwidth in H_z is therefore $f_0 = \frac{35300}{4\times 1.1} \approx 8000$. If the number of sections in the model is 16 (the length of the model is then $17.6cm$), there will be 8 peaks in the corresponding power spectrum, and the first three lowest of them are the formants F_1, F_2, F_3 needed for the study.

4.4 Mapping from Time to Articulatory Configurations – Articulatory Dynamics

This section discusses how to conduct the mapping from time to articulatory configurations of the model used for the simulation of sounds /aI/, /aU/ and /wi/. It includes 1) what parameters of the model are essentially varied for each sound, and 2) how each varying parameter changes with time,

What parameters are varied to simulate /aI/, /aU/ and /wi/?

It has been well known that location of the maximum constriction caused by the tongue body is a fundamental parameter of the vocal tract configuration for vowels (Stevens & House 1955, Fant 1960). In particular, Stevens' study on the quantal nature of speech based on the acoustic tube model has shown that there are three relatively stable regions along the vocal tract, where formant frequencies are relatively insensitive to the change of the maximum constriction place. These three regions correspond to three distinct classes of vowels: non-low front vowels (/i/-like vowel), non-low back vowels with lip closure (/u/-like vowel), and low vowels (/a/-like vowel) (Stevens 1972, 1989). It is obvious that the constriction location is an essential varying parameter for change from one of three classes of sounds to the other; that is, change of X_c is necessary for simulating /aI/, /aU/ and /wi/.

Kent (1972) and Kent & Moll (1972) reported some results of tongue movement during speech by observing several points marked on the tongue through cinefluorography. One of the results (Kent, 1972) is about the motion of the tongue body for the diphthong /aI/ in four different phonetic environments by one speaker. Three points on the tongue body were marked and observed. The two adjacent markers were connected by a straight line, and motion of the tongue body is represented by changes in the orientation of two straight line segments. It was found that during the articulation of diphthong /aI/, the line segments for this speaker "*move in a superior and slightly anterior direction*". The other result (Kent & Moll 1972) shows that for the two speakers in

pronouncing /aI/, the anterior tongue point moves both forward and upward.

The real vocal tract is more like a bent tube than a straight one. The high point of tongue body has two dimensional freedoms in the vocal tract. Kent's work suggests that the path of tongue body movement for /aI/ is more likely that displayed as in solid line than in dashed line in Figure 4-3. This indicates that simulation of /aI/ in the straight acoustic tube needs not only the changes of X_c , but also of R_c.

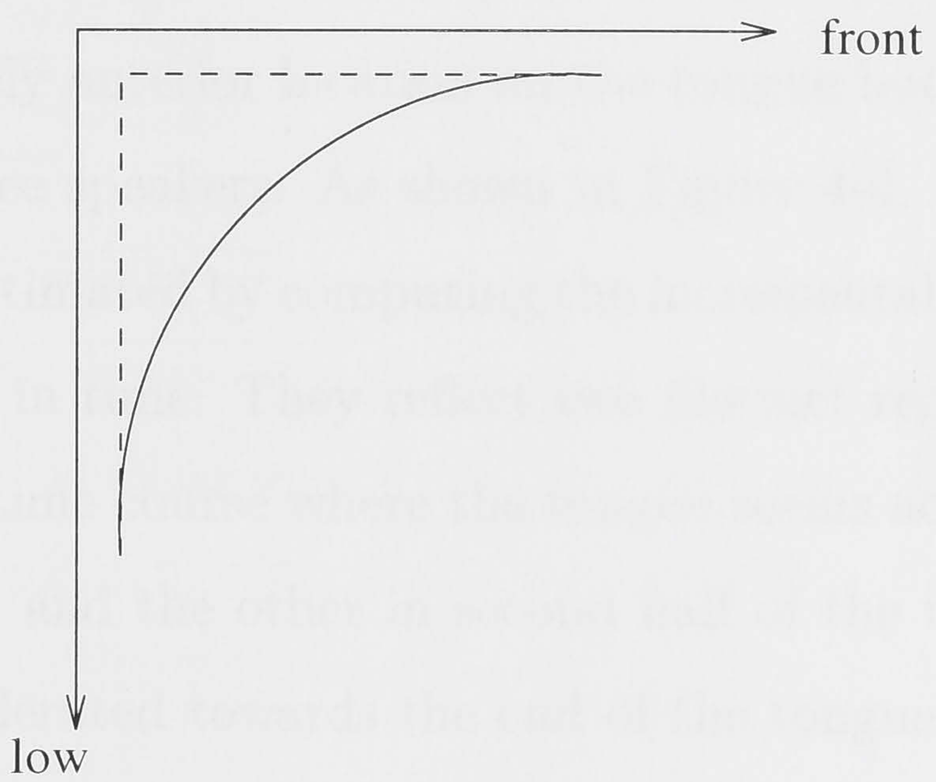


Figure 4-3: Illustration of possible path of tongue for /aI/.

For /aU/ and /wi/, lip rounding/unrounding is involved. In the acoustic tube model, change of R_l represents lip rounding/unrounding. It is known (e.g. Fant 1960) that lip rounding may lead to increase of the total vocal tract length. However in the model used for this study, the length of each section is fixed as $1.1cm$ and the total length of the model can not be changed continuously. Even if it could, this increase is small compared to the total length. Therefore, the model length is chosen to be fixed in the course of these two simulated sounds. In other words, we only simulate the realizations for /aU/ and /wi/ in which lip rounding does not cause increase of total length of the vocal tract.

In summary, Table 4.1 lists the parameters of the model needed to vary in simulation of /aI/, /aU/ and /wi/.

How does each varying parameter change with time?

Sounds	Time-varying parameters	
/aI/	$X_c(t)$	$A_c(t)(or R_c(t))$
/aU/	$X_c(t)$	$A_l(t)(or R_l(t))$
/wi/	$X_c(t)$	$A_l(t)(or R_l(t))$

Table 4.1: Time-varying parameters of the model in simulation of /aI/, /aU/ and /wi/.

Kent (1972) demonstrated the velocity-versus-time curves for the marked point having a relatively anterior location on the tongue body during the production of /aI/ by three speakers. As shown in Figure 4-4, those "triangular-shaped" curves were estimated by computing the incremental displacements for every $20ms$ increment in time. They reflect two distinct regions, one roughly in the first half of the time course where the tongue seems accelerated towards the maximum velocity and the other in second half of the time course where the tongue seems decelerated towards the end of the tongue movement.

Kent & Moll (1972) demonstrated three pairs of displacement-versus-time curves and velocity-versus-time curves for the anterior tongue point of one of two speakers during the vowel-to-vowel transitions in speech: a) /i/ to /ae/ in *free allies*; b) /i/ to /a/ in *he honoured*; c) /i/ to /or/ in *he ordered*. Three velocity-versus-time curves, measured in the same way as in (Kent 1972), were seen to have patterns similar to the velocity profile for /aI/ by three speakers reported by Kent (1972). The displacements were measured along the actual path of tongue-point movement and cumulated against time to produce the displacement-versus-time curves. These curves look like a sigmoidal function which reflects the time course of tongue movement between its initial and final positions.

Although Kent and Kent & Moll do not provide the displacement profiles for /aI/, /wi/ and /aU/, their results suggest that the high point of the tongue for these sounds might also have sigmoid-like characteristics in their displacement with time. Consequently in the experiments of this study, the

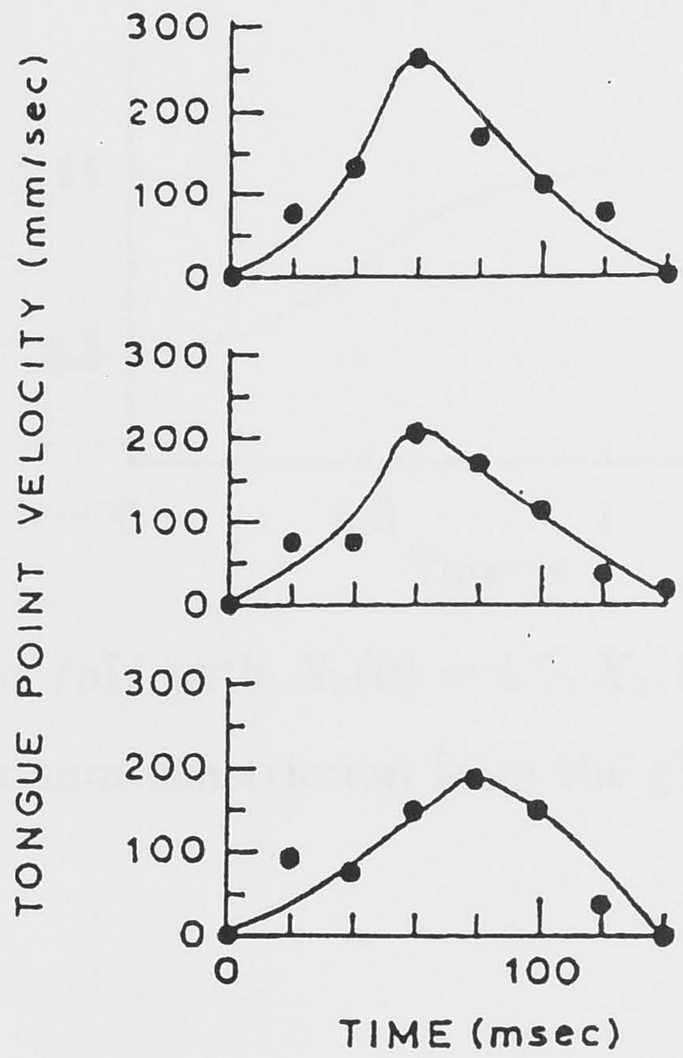


Figure 4-4: Velocity-versus-time curves for a point on the tongue body during the articulation of diphthong /aI/ by three speakers. the fleshpoint velocities were estimated by determining the incremental displacements for increments in time (20ms) (reprinted from Kent, 1972, p. 28).

constriction position as a function of time $X_c(t)$ for /aI/, /aU/ and /wi/ is assumed to take the form of sigmoidal function as follows, when the duration is scaled to unity.

$$X_c(t) = a + \frac{b}{1+e^{-c(t-d)}} \quad 0 \leq t \leq 1$$

where, $a = X_c(0) - \frac{b}{1+e^{c\times d}}$; $b = \frac{X_c(1)-X_c(0)}{p}$; $c = \frac{1}{d} \times \ln \frac{1+p}{1-p}$; $d = 0.5$; $p = 0.95$.

$X_c(0), X_c(1)$ are initial and final values of the maximum constriction position during producing each sound. p controls the sigmoidal characteristics of $X_c(t)$ and, therefore, determines its slope at different time points. The larger p is, the greater the difference between maximum and minimum slopes of $X_c(t)$. The value of 0.5 for d ensures the symmetrical function in time.

As an example, $X_c(t)$ for /aI/ is shown in Figure 4-5.

The constriction degree as a function of time $R_c(t)$ for /aI/ is also assumed to take a similar form

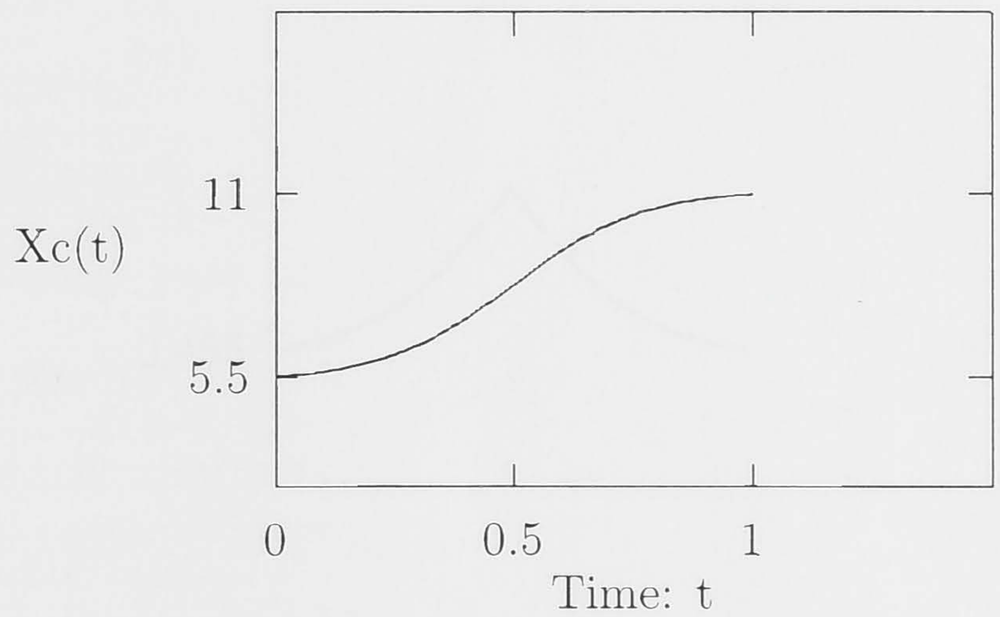


Figure 4-5: $X_c(t)$ for /aI/ with $X_c(0) = 5.5, X_c(1) = 11$. $X_c(t)$ represents distance of the maximum constriction from the glottis in cm.

$$R_c(t) = a + \frac{b}{1 + e^{-c(t-d)}}$$

But in order to simulate in the straight acoustic tube model, the path for that sound in the vocal tract, we have

for $0 \leq t \leq 0.5$:

$$a = R_c(0) - \tfrac{b}{1+e^{c\times d}};\ b = \tfrac{2(R_c(0.5)-R_c(0))}{p};\ c = \tfrac{1}{d} \times \ln \tfrac{1+p}{1-p};\ d = 0.5;\ p = 0.95.$$

for $0.5 \leq t \leq 1$:

$$a = 2(R_c(0.5) - R_c(1)) + R_c(1) - \tfrac{b}{1+e^{c\times d}};\ b = \tfrac{2(R_c(1)-R_c(0.5))}{p};\ c = \tfrac{1}{d} \times \ln \tfrac{1+p}{1-p};$$
$$d = 0.5;\ p = 0.95.$$

$R_c(0), R_c(0.5), R_c(1)$ are the values of constriction degree at $t = 0, 0.5, 1$ respectively.

The example of $R_c(t)$ for /aI/ is shown in Figure 4-6.

When demonstrating the velocity profile of the anterior tongue point for /aI/ by three speakers, Kent (1972) pointed out that "*Similar profiles have been reported for voluntary changes in the point of fixation of the eyes*". Then he commented that "*One might ask, then, if the triangular-shaped velocity profile is characteristic of certain types of controlled movements*". For lack

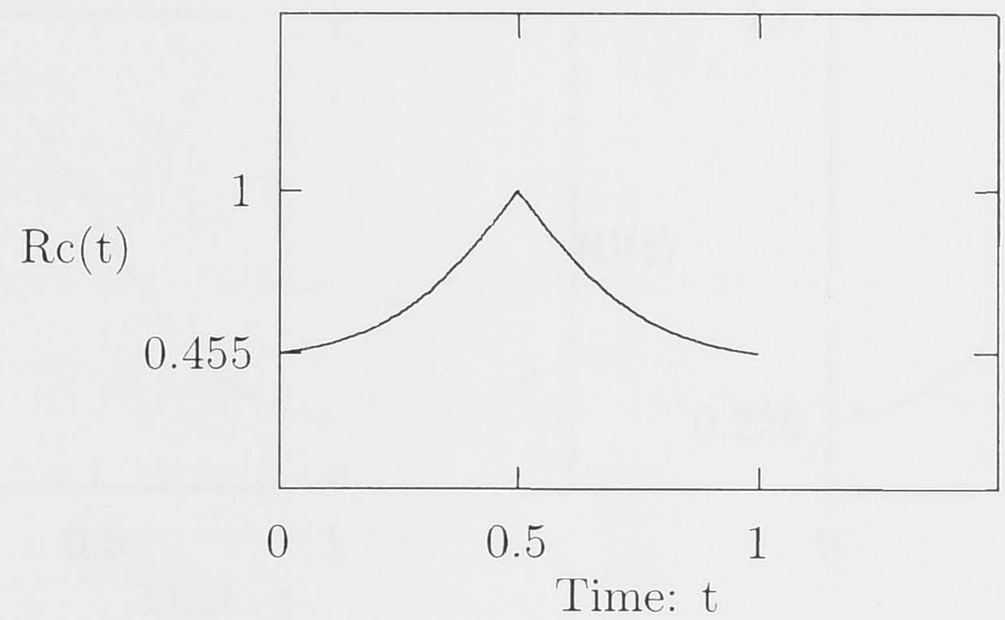


Figure 4-6: $R_c(t)$ for /aI/ with $R_c(0) = R_c(1) = .455, R_c(.5) = 1$. $R_c(t)$ represents radius of the cross-sectional area at the maximum constriction in cm.

of the knowledge about how lips move in rounding/unrounding, Kent's comments prompt me to assume that $R_l(t)$ in simulating /aU/ and /wi/ follows a sigmoidal function, which is identical to that assumed for $X_c(t)$,

$$R_l(t) = a + \frac{b}{1+e^{-c(t-d)}} \quad 0 \leq t \leq 1$$

where, $a = R_l(0) - \frac{b}{1+e^{c\times d}}$; $b = \frac{R_l(1)-R_l(0)}{p}$; $c = \frac{1}{d} \times \ln \frac{1+p}{1-p}$; $d = 0.5$; $p = 0.95$.

The curves of $R_l(t)$ from $t = 0$ to $t = 1$ are different between /aU/ and /wi/, because lips move from unrounding to rounding for /aU/ and vice versa for /wi/. This difference can be seen in Figure 4-7.

4.5 Boundary Conditions of Model Parameters for the Simulated Sounds

This section discusses choice of values of the model parameters which are associated with the articulatory structure in simulation of the three sounds.

In the experiments, the benchmark simulation for each sound was first performed. The values of the static model parameters are chosen either the same as or close to that in Fant's work (1960). For example, the length of the model L is chosen as 17.6 in the experiments rather than 17.5 as in Fant,

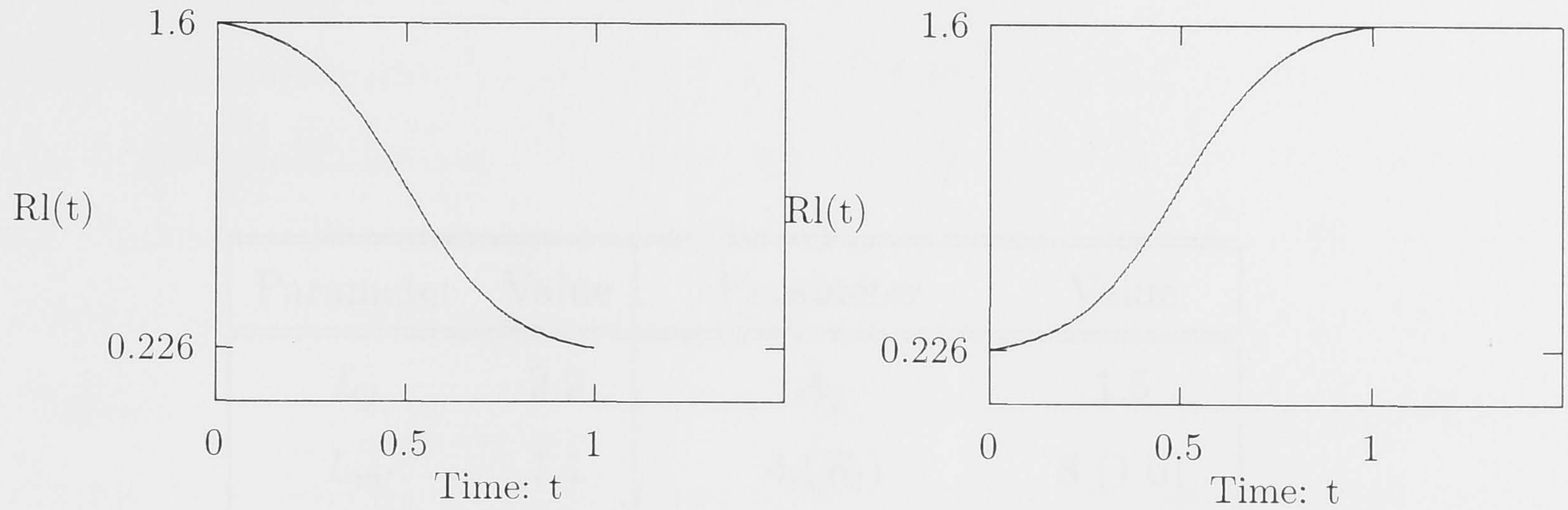


Figure 4-7: Left: $R_l(t)$ for /aU/ with $R_l(0) = 1.6, R_l(1) = 0.226$, right: $R_l(t)$ for /wi/ with $R_l(0) = .226, R_l(1) = 1.6$. $R_l(t)$ represents radius of the lips in cm.

because the length of a uniform section Δl was chosen as 1.1 and the number of sections as 16 in the acoustic tube model for this study.

The boundary values of the time-varying parameters for each sound are chosen basically according to that for the related static vowel sounds. The initial and final values of $X_c(t)$ for each sound are in the corresponding stable region of the nomogram derived from the model for this study. For sound /aI/, the choice of $R_c(0)$ $(= R_c(1))$ follows what Fant (1960) suggested for vowel /a/ and /i/, and $R_c(0.5)$ is arbitrarily chosen as one, larger than $R_c(0)$ $(= R_c(1))$ to simulate the change in this dimension. For /aU/ and /wi/, the boundary values of $R_l(t)$ $(A_l(t))$ are chosen also according to Fant. With lip rounding, it has the same value as that suggested for vowel /u/, and with lip unrounding, it has the same value as that suggested for vowel /a/ and /i/. Table 4.2 lists the relevant model parameters for the benchmark simulation of /aI/ in (a), /aU/ in (b) and /wi/ in (c).

Then, some parameters were perturbed to represent changes in size and shape of the vocal tract for the three sounds. It is assumed the length from $2.2cm$ to $5.5cm$ in the model represents the vertical pharyngeal portion of the vocal tract (as can be seen in Table 4.2, the larynx portion is assumed from $0cm - 2.2cm$, and $X_c(0)$ for the benchmark simulation of /aI/ and /aU/

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$A_l(R_l)$	8 (1.6)
L	17.6	$A_m(R_m)$	8 (1.6)
$X_c(0)$	5.5	$A_c(0)(R_c(0))$	.65 (.455)
$X_c(1)$	11	$A_c(0.5)(R_c(0.5))$	3.14 (1.)
		$A_c(1)(R_c(1))$	.65 (.455)

(a)

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$A_m(R_m)$	8 (1.6)
L	17.6	$A_c(R_c)$	.65 (.455)
$X_c(0)$	5.5	$A_l(0)(R_l(0))$	8 (1.6)
$X_c(1)$	8.5	$A_l(1)(R_l(1))$	.16 (.226)

(b)

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$A_c(R_c)$	.65 (.455)
L	17.6	$A_m(R_m)$	8 (1.6)
$X_c(0)$	9	$A_l(0)(R_l(0))$	.16(.226)
$X_c(1)$	12.5	$A_l(1)(R_l(1)$	8.(1.6)

(c)

Table 4.2: Values of the model parameters for the benchmark simulation of, /aI/ in (a), /aU/ in (b) and /wi/ in (c).

is $5.5cm$). If the perturbation in length occurs in that portion, the ratio of the length of the pharyngeal portion to the oral portion varies. It causes $X_c(0)$ and $X_c(1)$ for each sound to increase/decrease by the same amount if L increases/decreases by $1.1cm$ (the length of one section of the model). Equivalently, the length of the pharyngeal portion is increased by 1/3(33.3%). This seems realistic according to what Fant mentioned (1960, p.87) that "*the average male pharynx is 25 per cent longer than the average female pharynx, ...*". The perturbation also occurs in the maximum cross-sectional area and the cross-sectional area at the maximum constriction. To be consistent, their radius R_m and R_c are also increased/decreased by 33.3% of their values for the benchmark simulation. Although the rational for this perturbation choice of radius R_m and R_c is not very solid, I leave the corresponding simulated sounds to be verified later by listening. Eventually, total length L, maximum cross-sectional area A_m, cross-sectional area at the maximum constriction A_c (or the boundary values of $A_c(t)$ for sound /aI/ at $t = 0, 0.5, 1$) are perturbed from their benchmark values.

The model parameters under different conditions of the perturbed parameters for each sound are listed in Appendix A.

Under each condition in simulation of one sound, the temporal properties of varying parameters are held the same as discussed in Section 4.4, i.e. articulatory dynamics are kept identical.

4.6 Verification of Simulated Sounds by Listening

Each simulation of the sounds is synthesised in order to verify its phonetic quality by listening. With the articulatory dynamics discussed in Section 4.4 and the boundary conditions in Section 4.5 for each simulation, a sequence of 16th order reflection coefficients (when the length of model L is 17.6) can be

derived from the model. The program LP_SYN in ESPS is used to perform speech waveform synthesis from the reflection coefficients and the parametric source.

A file for the parametric source contains the information of 1) the RMS value, 2) the fundamental frequency, 3) the voicing state, for each frame. The values are chosen as 5000, 100 and 1 respectively. These values are the same for all frames within the parametric source file; this means the source property is not time-varying within one sound. The parametric source file is used for all different simulations of each of three sounds.

I first listened to the synthesised speech waveforms of the benchmark simulation of /aI/ as in "high", /aU/ as in "how" and /wi/ as in "we". They sounded like what I expected in phonetic quality and as produced by a male speaker. Then, for each sound, I listened to those of the simulations with variation of total length L, maximum cross-sectional area A, cross-sectional area at the maximum constriction A_c, which have been described in the previous section. The phonetic quality of each simulation was identified the same as that of its corresponding benchmark simulation. For some simulations, differences in auditory quality from their benchmark counterpart can be recognised, but for the others, they can not be perceived. Note that the differences perceived by listening are due to the variation in static structure of the model rather than that in its dynamics.

4.7 Experimental Results: The Acoustic Dynamics

The formant trajectories as a function of time for each simulation of a sound are derived by performing two successive mappings: 1) time is uniformly increased from 0 to 1; a pair of time varying parameters (e.g. $X_c(t), R_c(t)$ for /aI/) for each t are collected according to the articulatory dynamics discussed in

the Section 4.4, which together with other static parameters determines the model configuration at t, 2) given the configuration of the model at t, the corresponding first three formant values $F_1(t), F_2(t), F_3(t)$ are obtained using the method described in Section 4.3.

The detail experimental results for each sound simulated under different static conditions are reported as follows.

4.7.1 Tongue Moving from Low-back to High-front

$F_1(t), F_2(t), F_3(t)$ for the benchmark simulation of /aI/ are shown as curves (b) in Figure 4-8. The patterns look similar to part of the vocalic segment of "high"s in the Monosyllabic Words database described in Chapter 2 (one of them, as an example is shown in Figure 4-9). Especially, the V-shaped pattern of $F_3(t)$, which was also observed by a previous study on diphthong /aI/ (Clermont, 1991, 1993), has been realised. The trajectories of the first three formants of the acoustic tube model with the same conditions as for the benchmark simulation but $A_c(t)$ fixed as constant ($= 0.65cm^2$) throughout the duration are presented in Figure 4-8, labelled as curves (a). Curves(a) are similar to what Stevens & House (1955), Fant (1960) and Stevens (1972, 1989) reported in the equivalent region of their acoustic tube models on how formants change as the constriction moves forward. The differences between curves (a) and (b) can be illustrated in terms of the slopes of $F_1(t)$ and $F_2(t)$ and the pattern of $F_3(t)$. These differences indicate that, it is the variation in constriction degree during the advancement of constriction position that causes the changes in the slopes of $F_1(t), F_2(t)$ and the V-shape of $F_3(t)$.

The first three formant trajectories with variation of the acoustic tube length are shown in Figure 4-10. The relevant parameters chosen for these trajectories are listed in Table A.1.

As can be seen in Figure 4-10, the formant trajectories generally move down as L increases. This is consistent with what would be predicted by acoustic theory (Fant, 1960) and was verified by Stevens & House (1963) measuring

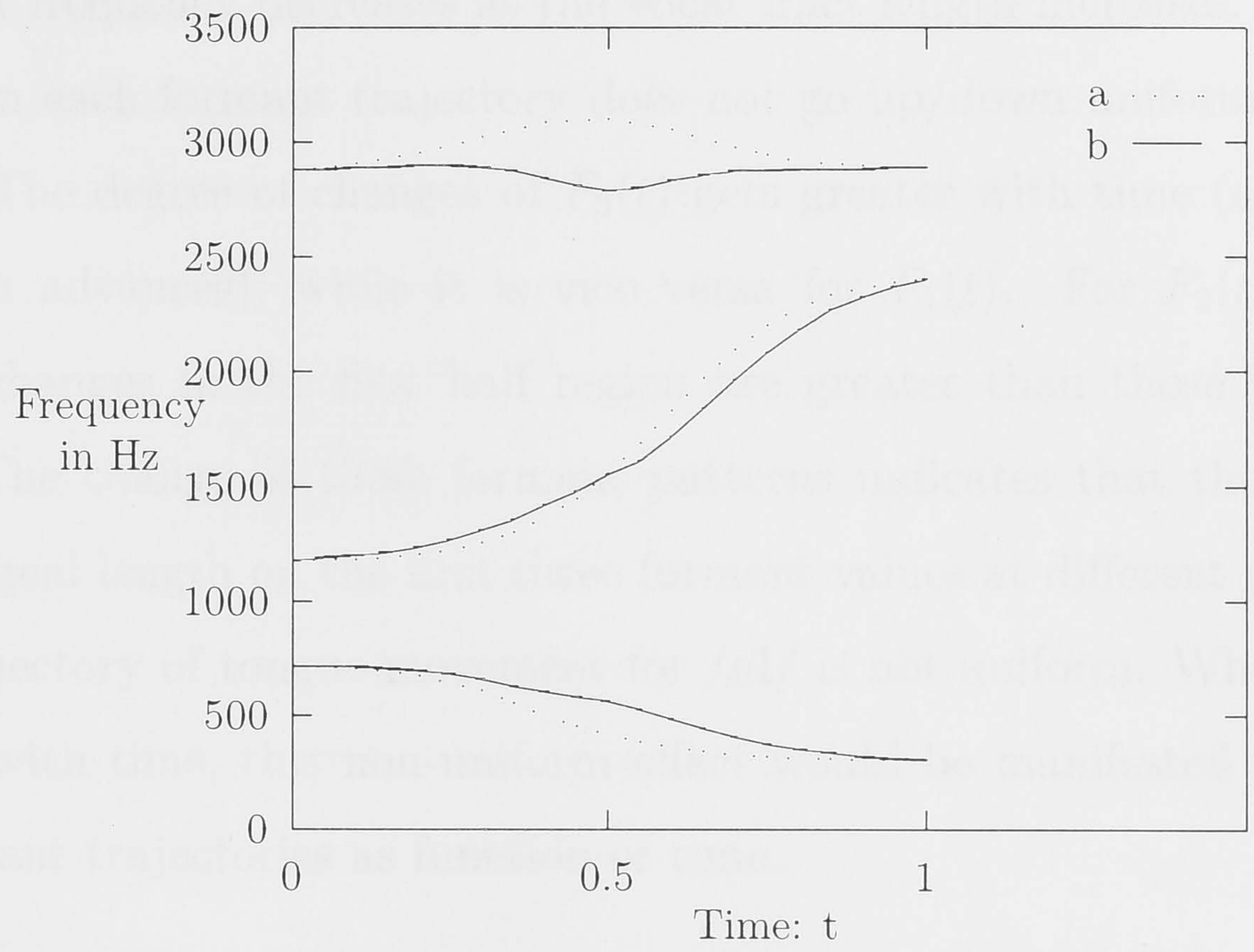


Figure 4-8: The first three formant trajectories, (b): benchmark simulation of /aI/. (a): same in model parameters as (b) but $A_c(t)$ fixed as 0.65.

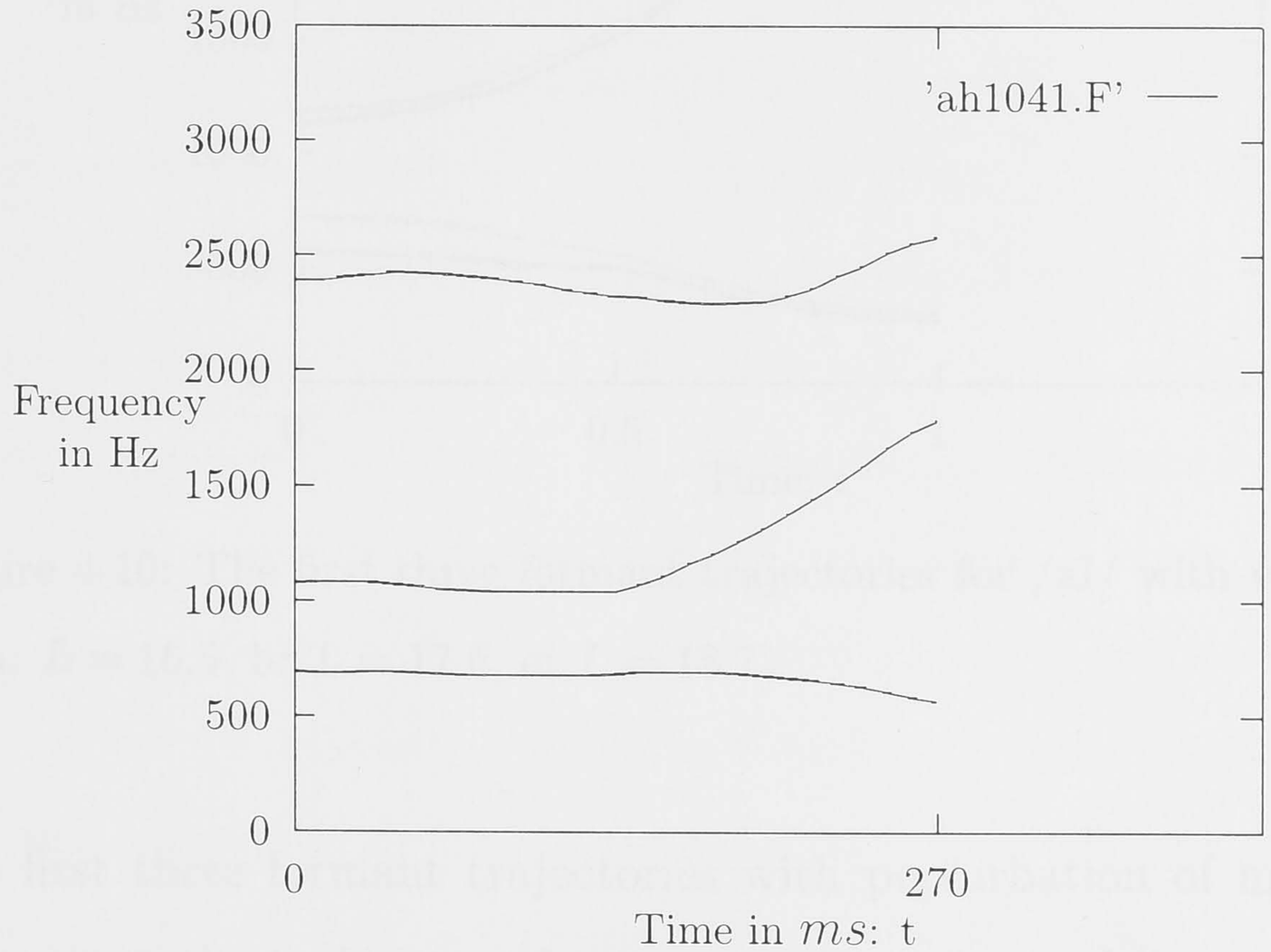


Figure 4-9: The first three formant trajectories of /aI/ in a "high", spoken by speaker AH.

the vocal tract length from X-ray pictures of several vowels for three speakers: formant frequency decreases as the vocal tract length increases. However, the value on each formant trajectory does not go up/down uniformly along time scale. The degree of changes of $F_3(t)$ gets greater with time (as constriction position advances), while it is vice versa for $F_1(t)$. For $F_2(t)$, it looks as if the changes in the first half region are greater than those in the second half. The change of these formant patterns indicates that the effect of the pharyngeal length on the first three formant values at different position along the trajectory of tongue movement for /aI/ is not uniform. When the tongue moves with time, this non-uniform effect would be manifested in the change of formant trajectories as function of time.

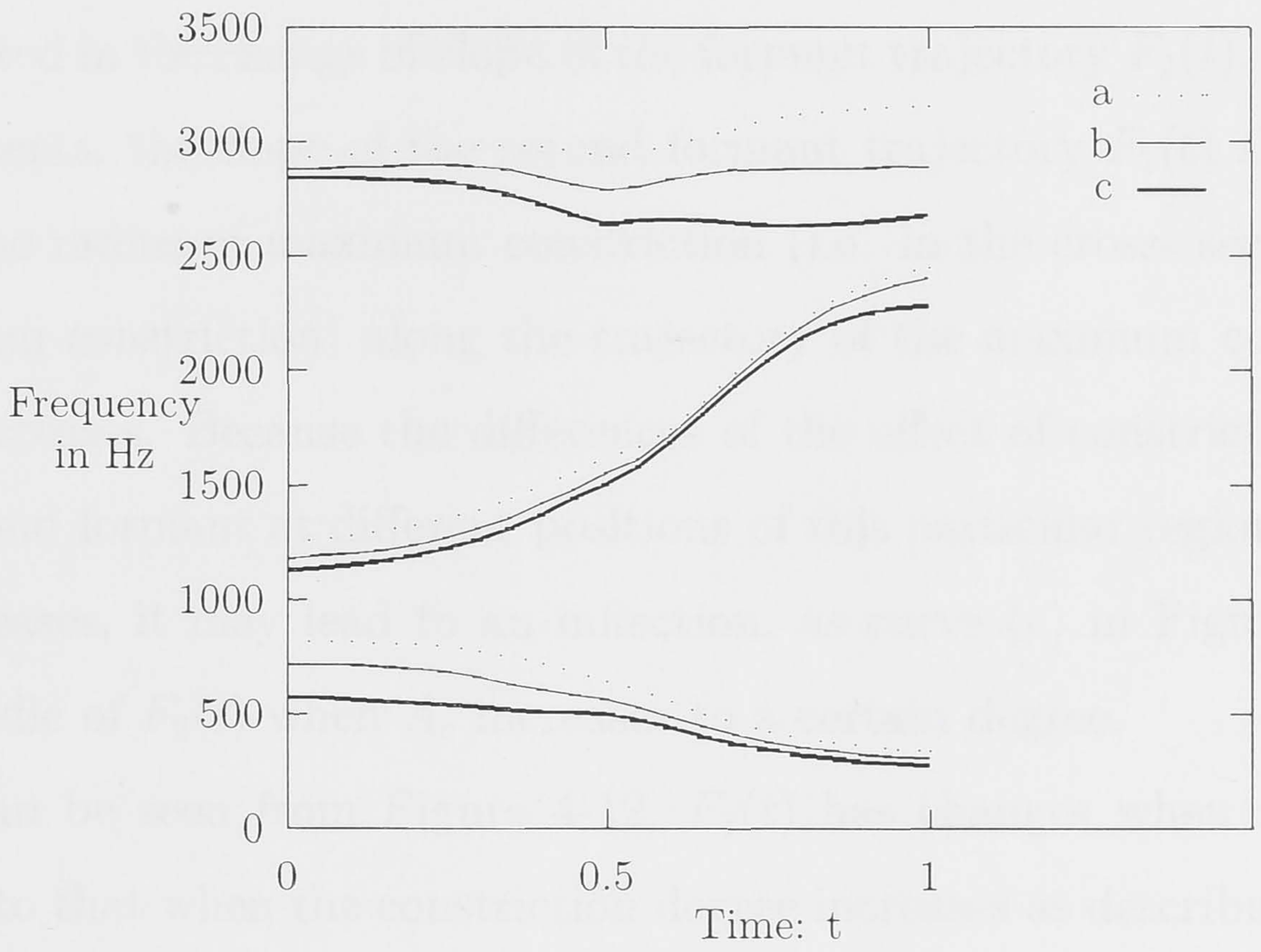


Figure 4-10: The first three formant trajectories for /aI/ with variation of L, a: $L = 16.5$, b: $L = 17.6$, c: $L = 18.7$.

The first three formant trajectories with perturbation of maximum constriction along the trajectory of tongue movement are shown in Figure 4-11. And the first three formant trajectories with perturbation of the maximum cross-sectional area are shown in Figure 4-12. The relevant parameters are listed in Table A.2 and Table A.3, respectively.

From Figure 4-11, it is observed that with increased constriction along the maximum constriction trajectory for /aI/, $F_1(t)$ shifts downwards in a way that is monotonically related to the degree of constriction; $F_2(t)$ has opposite effects for the first and second half region and hence the slope of $F_2(t)$ increases; $F_3(t)$ increases and so does the concavity of its V-shape.

In the equivalent region of X_c, the similar phenomena have been observed by Stevens & House (1955) and Fant (1960) when they studied the effects of constriction degree on the formants at different constriction positions in the acoustic tube models. When describing F_2, Stevens & House noticed that "*the rate of increase* (note: it refers to the increase of F_2 as the constriction moves forward) *depending markedly on the size of the constriction*". When the constriction moves forward with time, this effect of constriction degree will be manifested in the change of slope of the formant trajectory $F_2(t)$. Hence, in the experiments, the slope of the second formant trajectory $F_2(t)$ shows increase when the radius of maximum constriction (i.e. in the cross- sectional area of maximum constriction) along the trajectory of the maximum constriction for /aI/ decreases. Because the differences of the effect of constriction degree on the second formant at different positions of this particular region decreases as A_c increases, it may lead to an inflection, as curve (c) in Figure 4-11 round the middle of $F_2(t)$ when A_c increases to a certain degree.

It can be seen from Figure 4-12, $F_3(t)$ has changes when A_m increases, similar to that when the constriction degree increases as described above. The decrease of F_1, and the increase of $F_2(t)$ with time, is more pronounced as A_m increases. These are because the maximum cross sectional area affects the formant values differently along the tongue movement trajectory for /aI/. As a result, there is a change in formant patterns as function of maximum constriction position of the tongue. As the tongue moves with time, this change will be shown in the change of the formant patterns as function of time. An inflection in middle of $F_2(t)$ is also seen in curves (c) of Figure 4-12. This is because the differences of the effect of maximum cross area on the

second formant at different positions of this particular region decreases as A_m decreases.

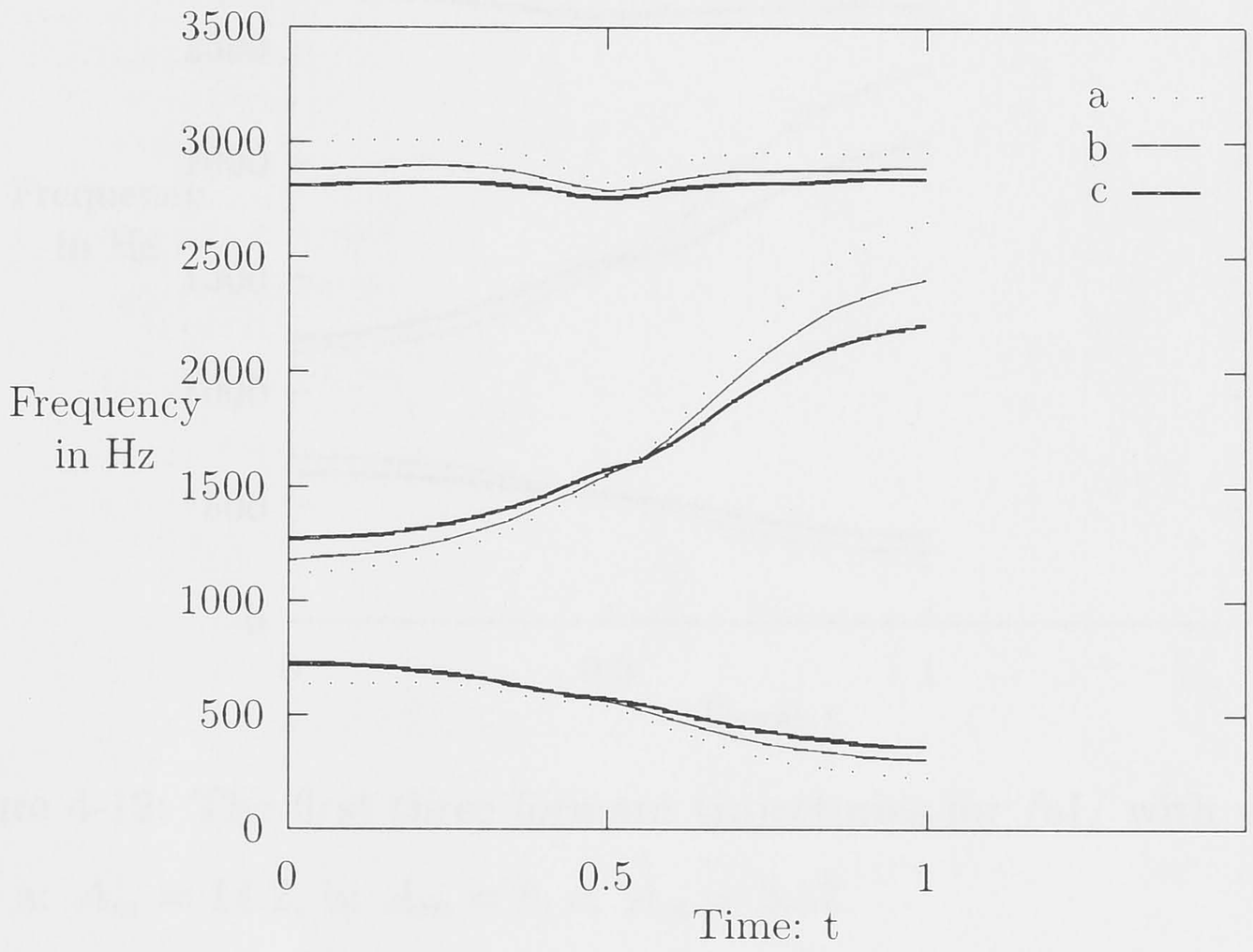


Figure 4-11: The first three formant trajectories for /aI/ with variation of $A_c(0)$; a: $A_c(0) = .289$, b: $A_c(0) = .65$, c: $A_c(0) = 1.16$.

4.7.2 Tongue Moving from Low-back to High-back with Lips from Unrounding to Rounding

$F_1(t), F_2(t), F_3(t)$ for the benchmark simulation of /aU/ are shown in Figure 4-13 as curves (b). The patterns of (b) look similar to part of the vocalic segment of "how"s in the Monosyllabic Words database described in Chapter 2 (one of them as an example is shown in Figure 4-14). The trajectories of the first three formants of the acoustic tube model with the same conditions as for the benchmark simulation but $A_l(t)$ fixed as constant ($= 8cm^2$) throughout the duration are also presented in Figure 4-13, labelled as curves (a). Curves(a) are similar to what Stevens & House (1955), Fant (1960) and Stevens (1972, 1989) reported in the equivalent region of their acoustic tube models on how formants change as the constriction moves forward. The values on $F_1(t), F_2(t), F_3(t)$

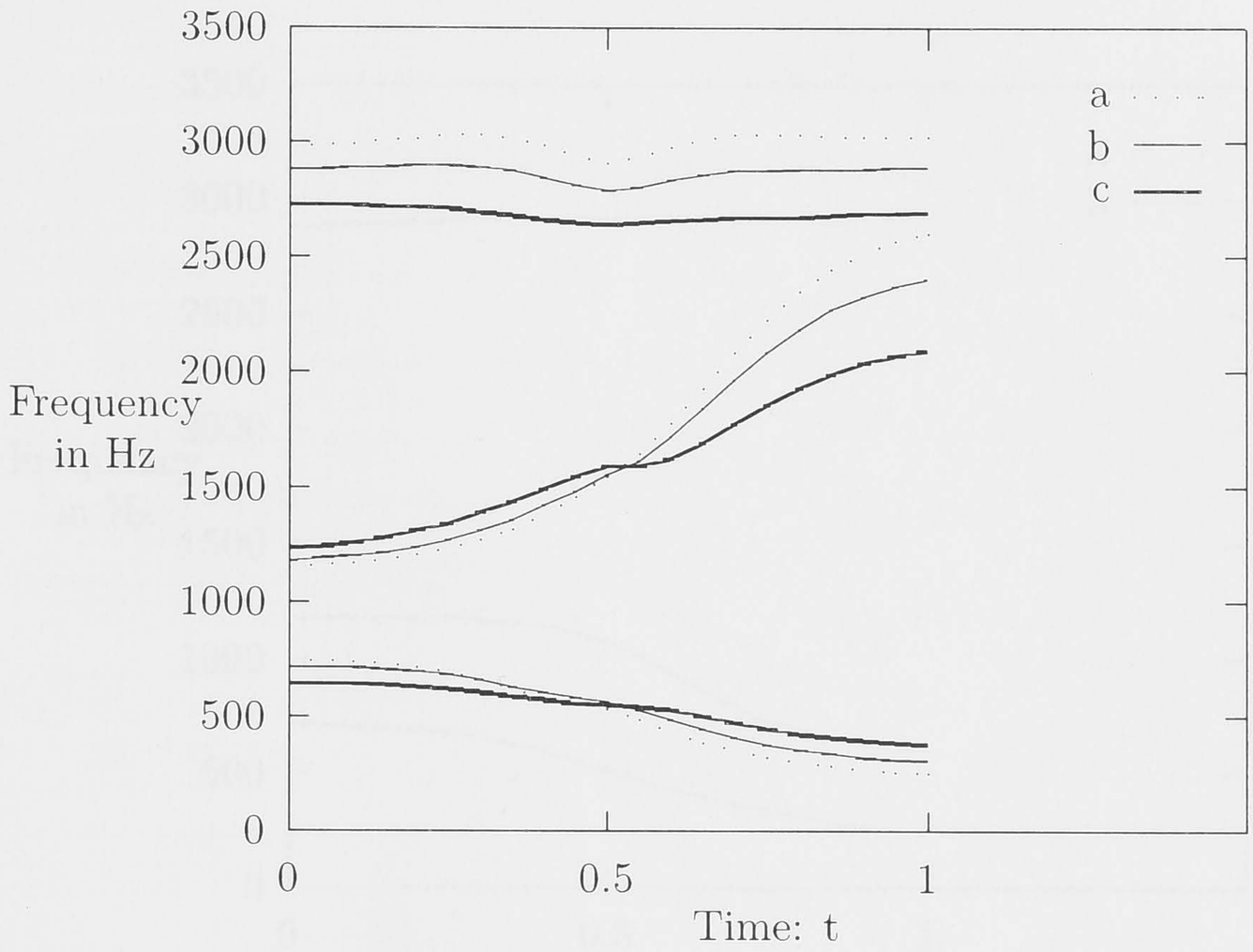


Figure 4-12: The first three formant trajectories for /aI/ with variation of A_m; a: $A_m = 14.2$, b: $A_m = 8$, c: $A_m = 3.57$.

shown in curves (b) are falling down towards the end of trajectories due to decrease of $R_l(t)$. Particularly, the decrease of $R_l(t)$ makes F_2 monotonically decrease rather than increase as when R_l is held constant. Figure 4-13 indicates that lips movement from unrounding to rounding is another predominant factor for $F_1(t), F_2(t), F_3(t)$ patterns of /aU/.

The first three formant trajectories with variation of the acoustic tube length are shown in Figure 4-15. The relevant parameters chosen for these trajectories are listed in Table A.4. As expected, the formant trajectories move down as L increases. The contours of $F_2(t), F_3(t)$ do not show much sensitive to L in terms of the absolute values and the non-uniform change over the trajectory. However it is clear that $F_1(t)$ is more sensitive, reflecting that the effect of the pharyngeal length on the F_1 decreases as the articulators (tongue and lips) move towards the end of their movement trajectories for /aU/.

The first three formant trajectories with variation of constriction degree are shown in Figure 4-16. And the first three formant trajectories with variation

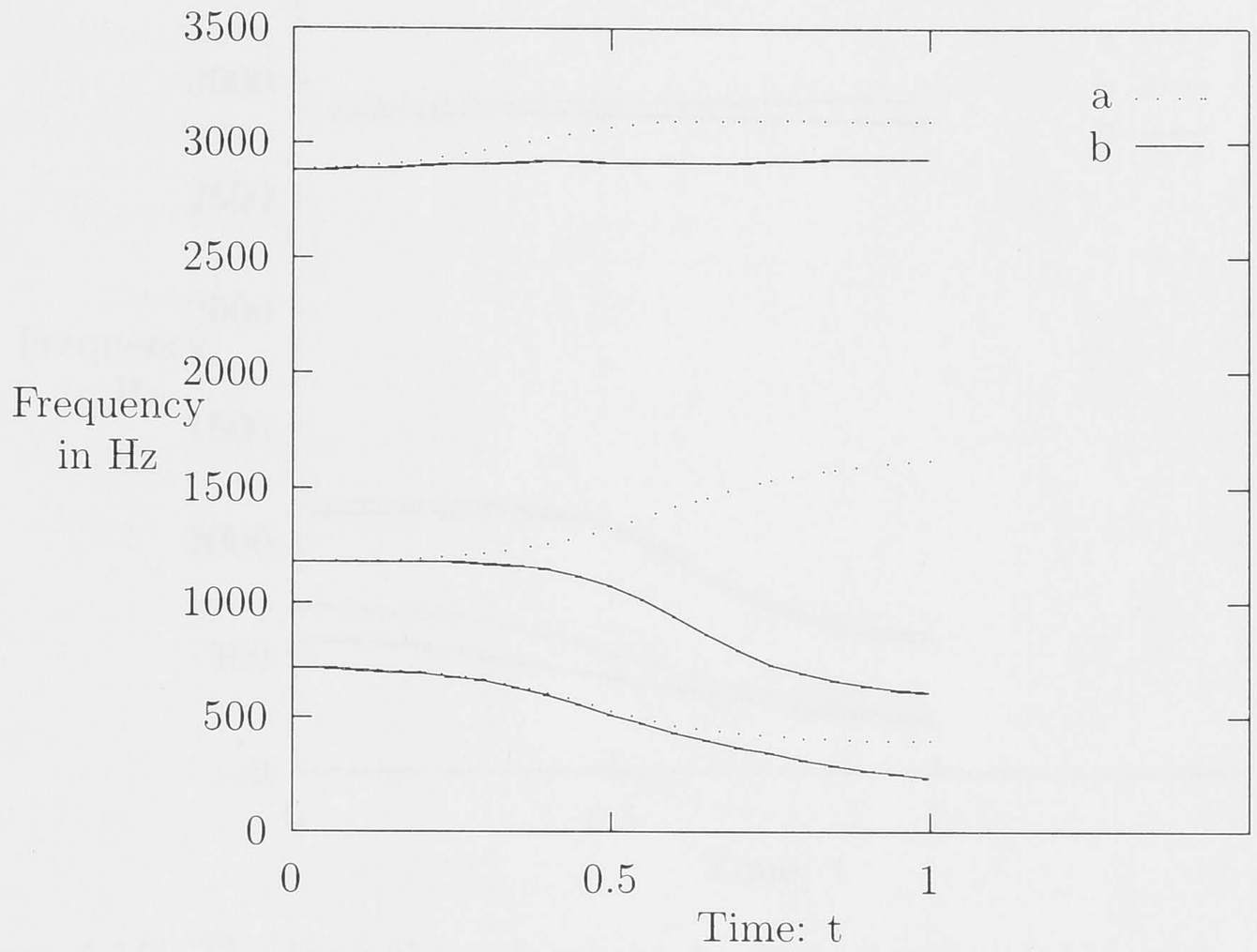


Figure 4-13: Simulated first three formant trajectories, b: benchmark simulation for /aU/. (a): same in model parameters as (b) but $A_l(t)$ fixed as 8.

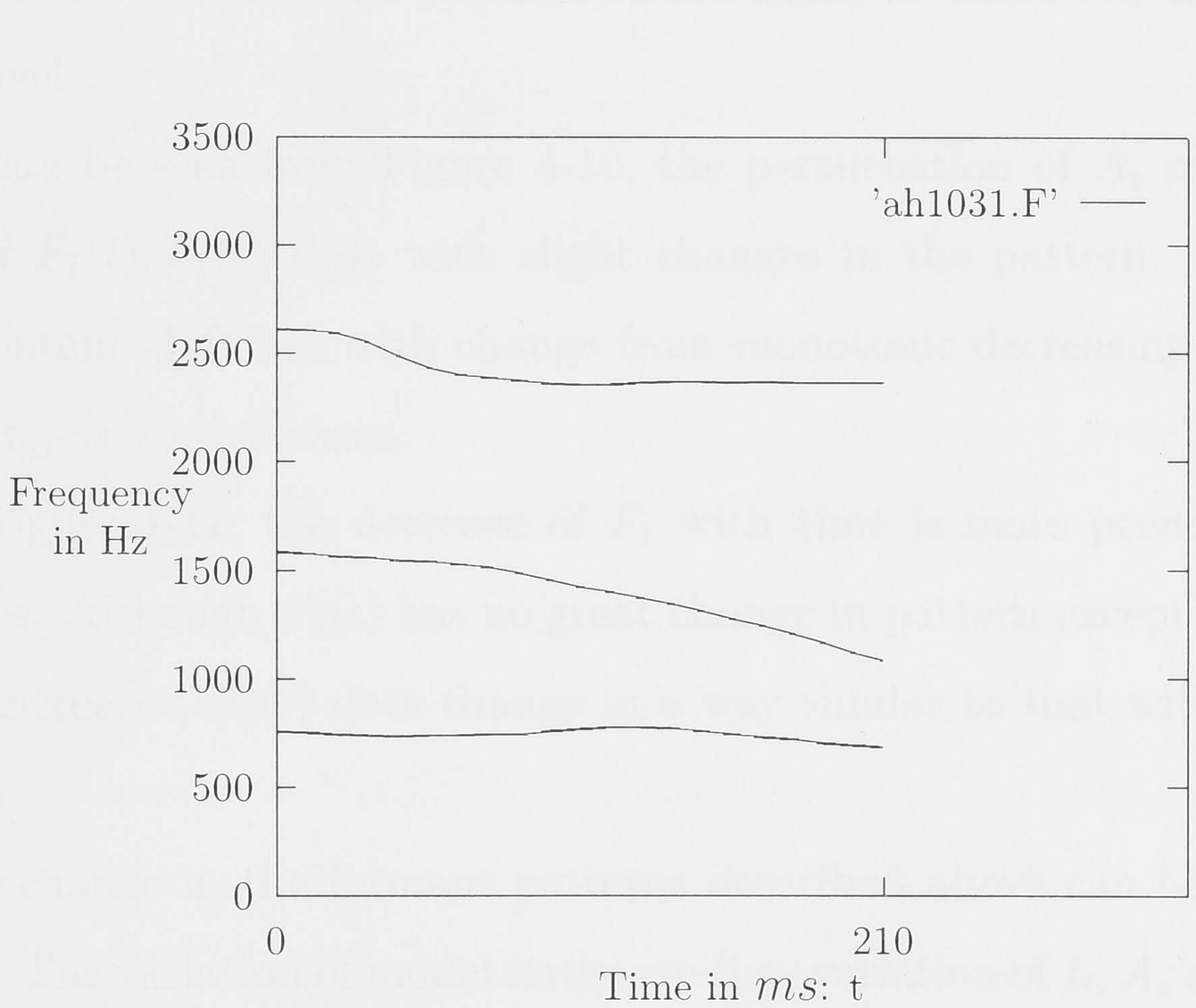


Figure 4-14: The first three formant trajectories of /aU/ in a "how", spoken by speaker AH.

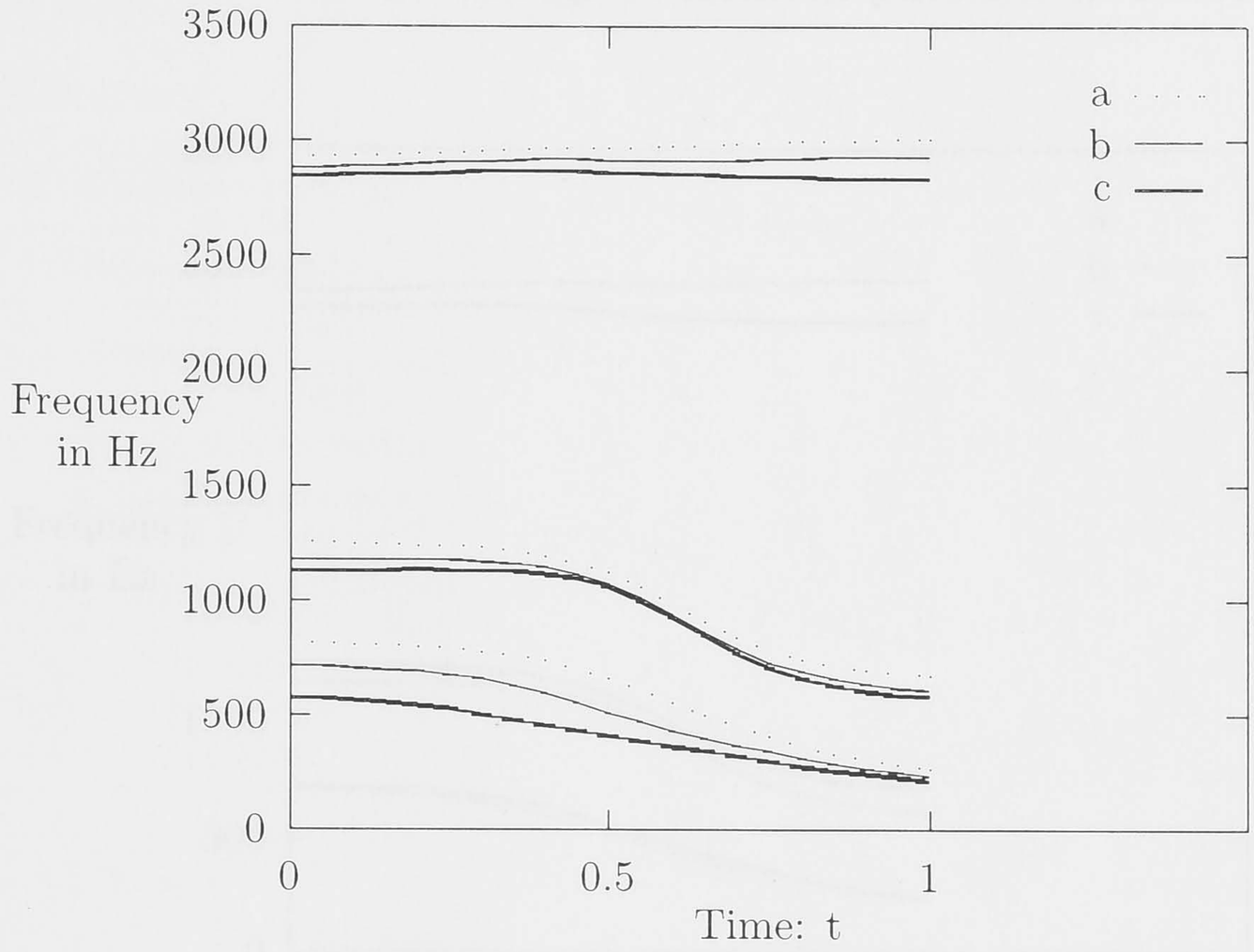


Figure 4-15: The first three formant trajectories for /aU/ with variation of L, a: $L = 16.5$, b: $L = 17.6$, c: $L = 18.7$.

of the maximum cross-sectional area of the acoustic tube model are shown in Figure 4-17. The relevant parameters are listed in Table A.5 and Table A.6, respectively.

As can be seen from Figure 4-16, the perturbation of A_c makes the contours of $F_1(t), F_2(t)$ shift with slight changes in the pattern. It also makes $F_3(t)$ contour shift but with change from monotonic decreasing to monotonic increasing as A_c decreases.

In Figure 4-17, the decrease of F_1 with time is more pronounced as A_m increases. Although $F_2(t)$ has no great change in pattern except shifting down as A_m increases, $F_3(t)$ does change in a way similar to that with decreases of A_c.

The change in the formant patterns described above can be explained as follows. The variation of model structure (i.e. variation of L, A_c and A_m) could affect the formants differently at the different positions along the trajectory of tongue and lips movement for /aU/. When these articulators move with time, the change of the formant patterns as function of the tongue position and lips

size will be observed in the change of formant patterns as function of time.

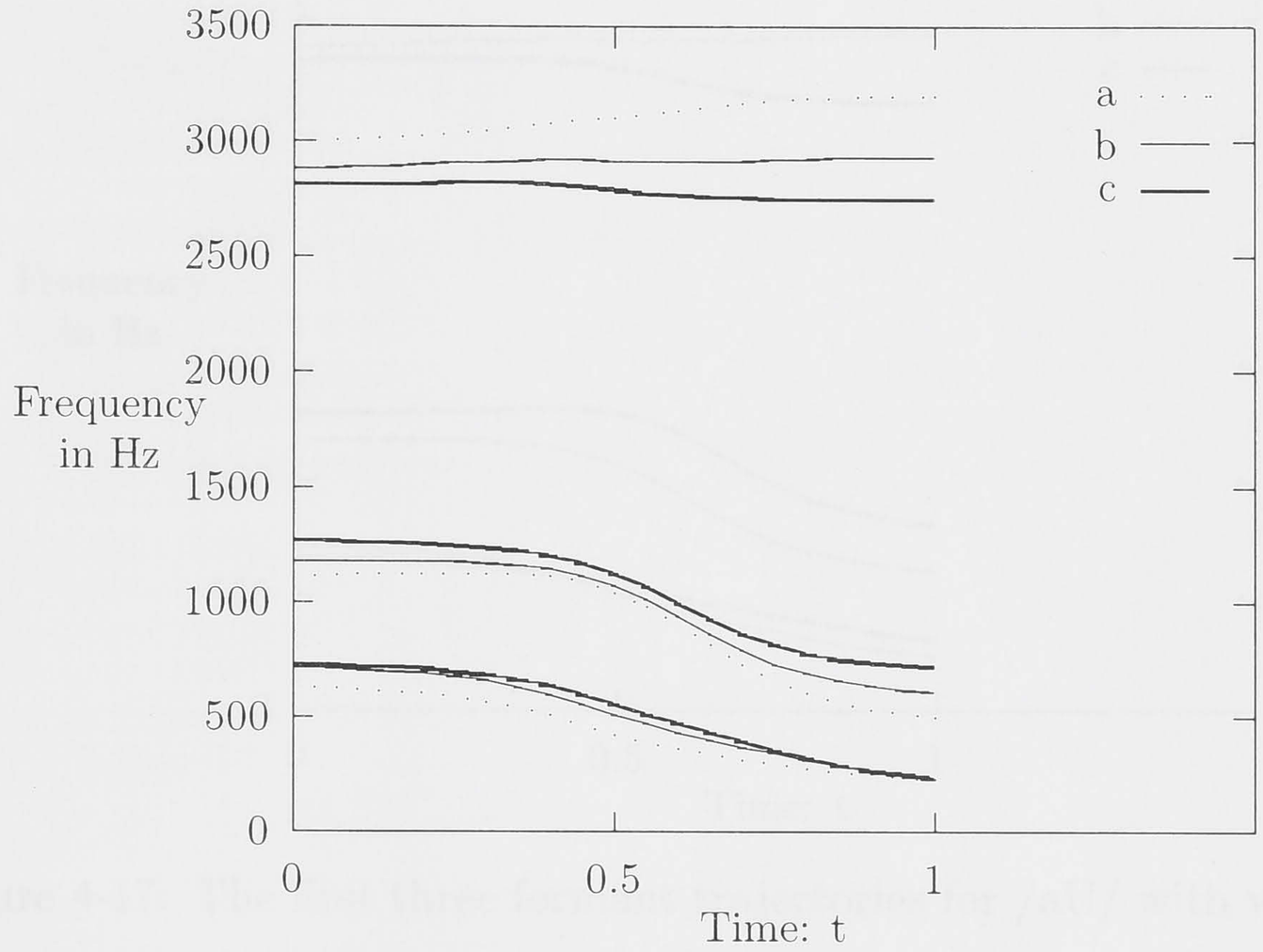


Figure 4-16: The first three formant trajectories for /aU/ with variation of A_c, a: $A_c = .289$, b: $A_c = .65$, c: $A_c = 1.16$.

4.7.3 Tongue Moving from High-back to High-front with Lips from Rounding to Unrounding

$F_1(t), F_2(t), F_3(t)$ for the benchmark simulation of /wi/ are shown as curves (b) in Figure 4-18. The patterns of curves (b) look similar to part of the vocalic segment of "we"s in the Monosyllabic Words database described in Chapter 2 (one of them as an example is shown in Figure 4-19). The trajectories of the first three formants from the model with the same conditions as for the benchmark simulation but $A_l(t)$ fixed as constant ($= 8cm^2$) throughout the duration are presented in Figure 4-18, labelled as curves (a). The patterns of curves (a) are similar to what Fant (1960) and Stevens (1972, 1989) reported on the formant patterns as a function of constriction formed in the anterior part of the model. It can be observed that the location of the maximum in F_2 is very near to the location of the minimum in F_3, while F_1 decreases

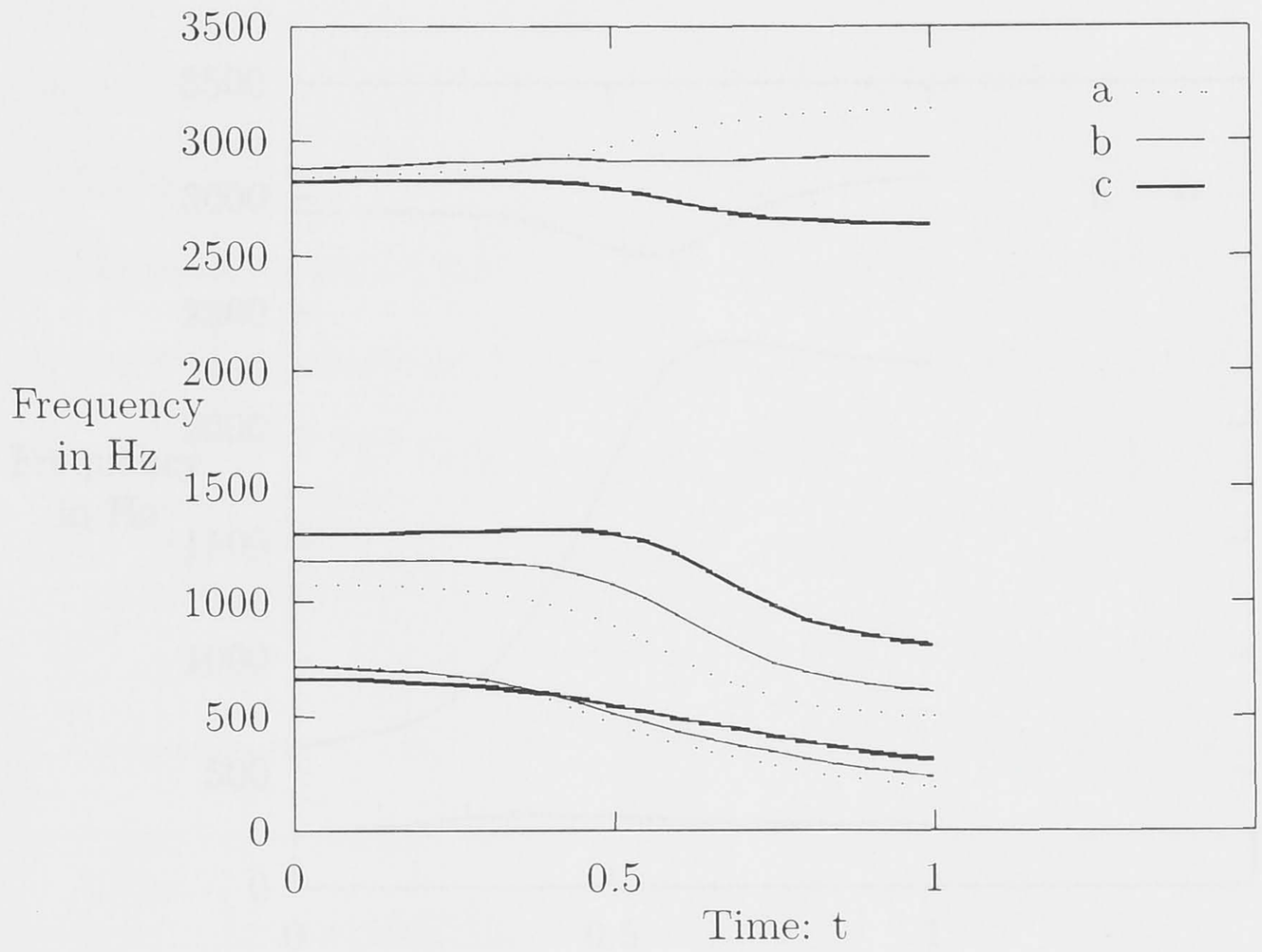


Figure 4-17: The first three formant trajectories for /aU/ with variation of A_m, a: $A_m = 14.2$, b: $A_m = 8$, c: $A_m = 3.57$.

monotonically. All F_1, F_2, F_3 are reduced at the beginning of the trajectories due to decrease of $A_l(t)$. Particularly, the decrease of A_l makes $F_2(0)$ very strongly decrease, and consequently makes the slope increase. Figure 4-18 indicates that lip movement from rounding to unrounding is a predominant factor for the patterns of $F_1(t), F_2(t), F_3(t)$ of /wi/.

The first three formant trajectories with variation of the acoustic tube length are shown in Figure 4-20. The relevant parameters chosen for these trajectories are listed in Table A.7. The increase of the pharyngeal length moves $F_1(t)$ down. It shifts the offset region of $F_2(t)$ down, and so does it with $F_3(t)$ but increases the V-shape concavity of $F_3(t)$. As a results, it changes the $F_2(t)$ and $F_3(t)$ pattern with decrease of their minimum distance.

The first three formant trajectories with variation of the maximum cross-sectional area of the acoustic tube model are shown in Figure 4-21. And the first three formant trajectories with variation of constriction degree are shown in Figure 4-22. The relevant parameters are listed in Table A.8 and Table A.9, respectively.

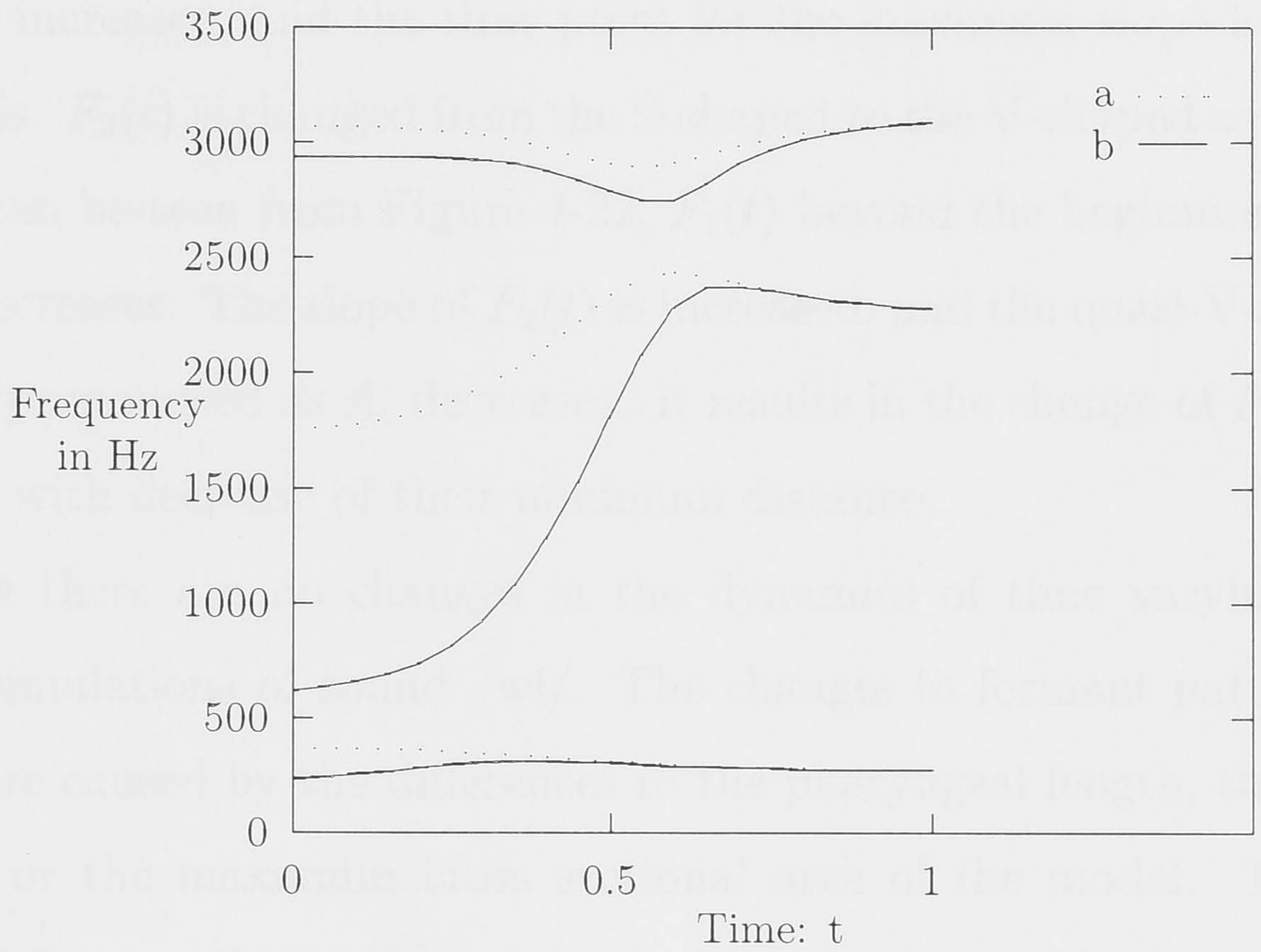


Figure 4-18: Simulated first three formant trajectories, b: benchmark simulation for /wi/. a: same in model parameters as (b) but $A_l(t)$ fixed as 8.

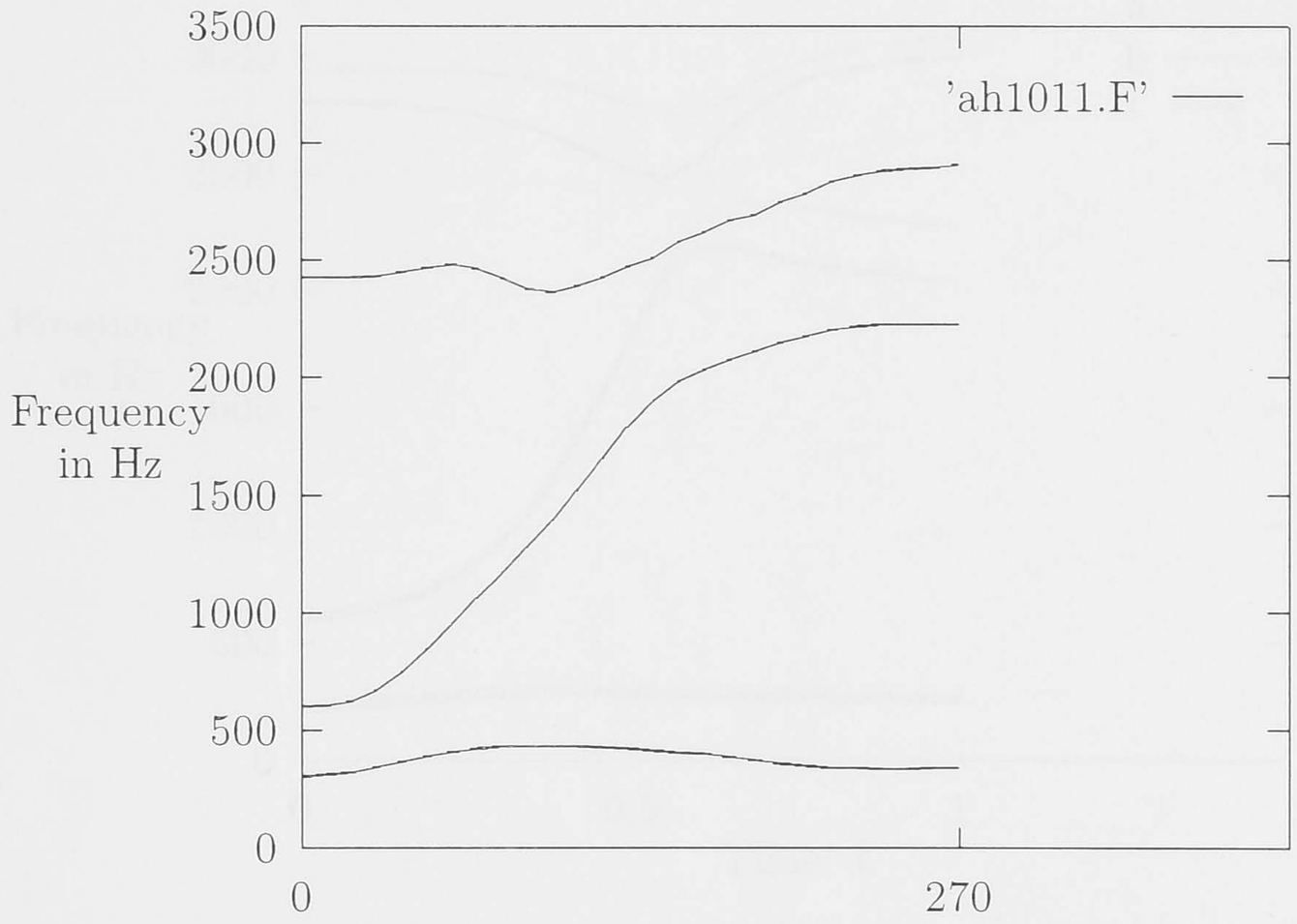


Figure 4-19: The first three formant trajectories of /we/ in a "we", spoken by speaker AH.

Figure 4-21 shows that, $F_1(t)$ moves down as A_m increases, The slope of $F_2(t)$ is increased, and the time point for the maximum slope is delayed as A increases. $F_3(t)$ is changed from the S-shaped to the V-shaped as A_m increases.

As can be seen from Figure 4-22, $F_1(t)$ beyond the beginning moves down as A_c decreases. The slope of $F_2(t)$ is increased, and the quasi-V-shape of $F_3(t)$ is more pronounced as A_c decreases. It results in the change of $F_2(t)$ and $F_3(t)$ pattern with decrease of their minimum distance.

Note there are no changes in the dynamics of time varying parameters for all simulations of sound /wi/. The changes in formant pattern described above are caused by the differences in the pharyngeal length, the constriction degree, or the maximum cross sectional area of the model. These changes reflect different effects of the perturbed parameters on the formants along the trajectories of the tongue and lips movement for /wi/. As the articulators move with time, those effects are surely manifested in the change of formant dynamics.

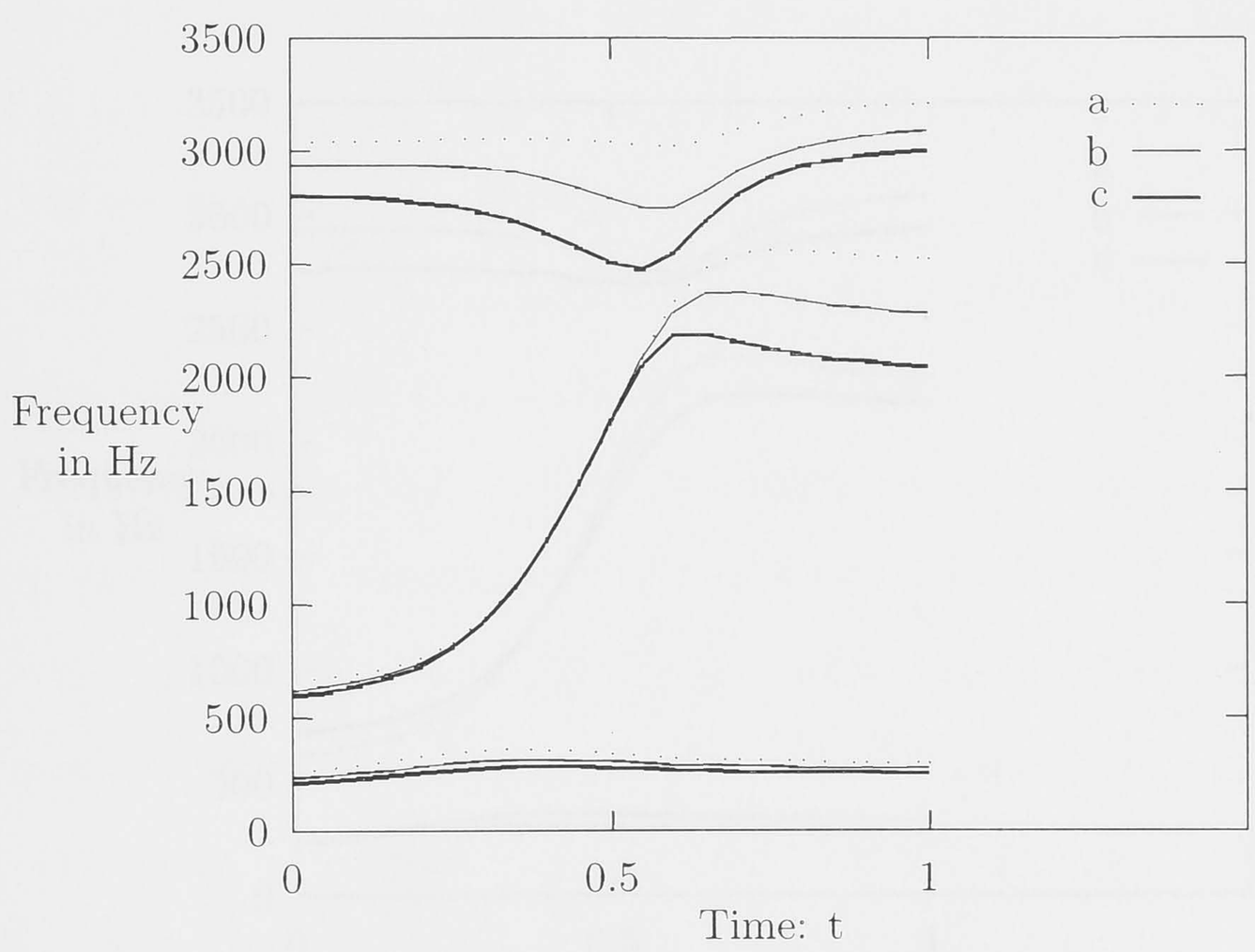


Figure 4-20: The first three formant trajectories for /wi/ with variation of L, a: $L = 16.5$, b: $L = 17.6$, c: $L = 18.7$.

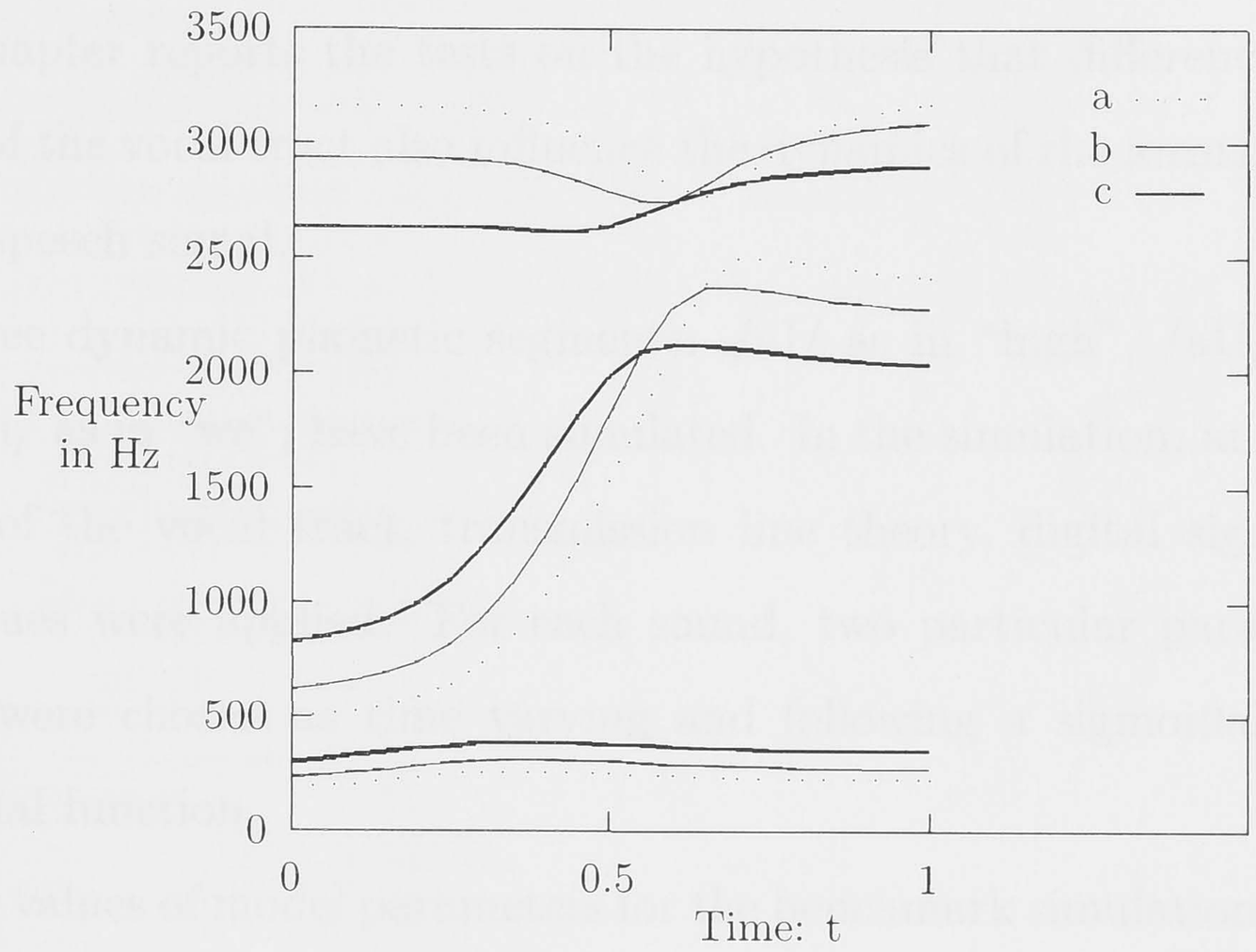


Figure 4-21: The first three formant trajectories for /wi/ with variation of A_m, a: $A_m = 14.2$, b: $A_m = 8$, c: $A_m = 3.57$.

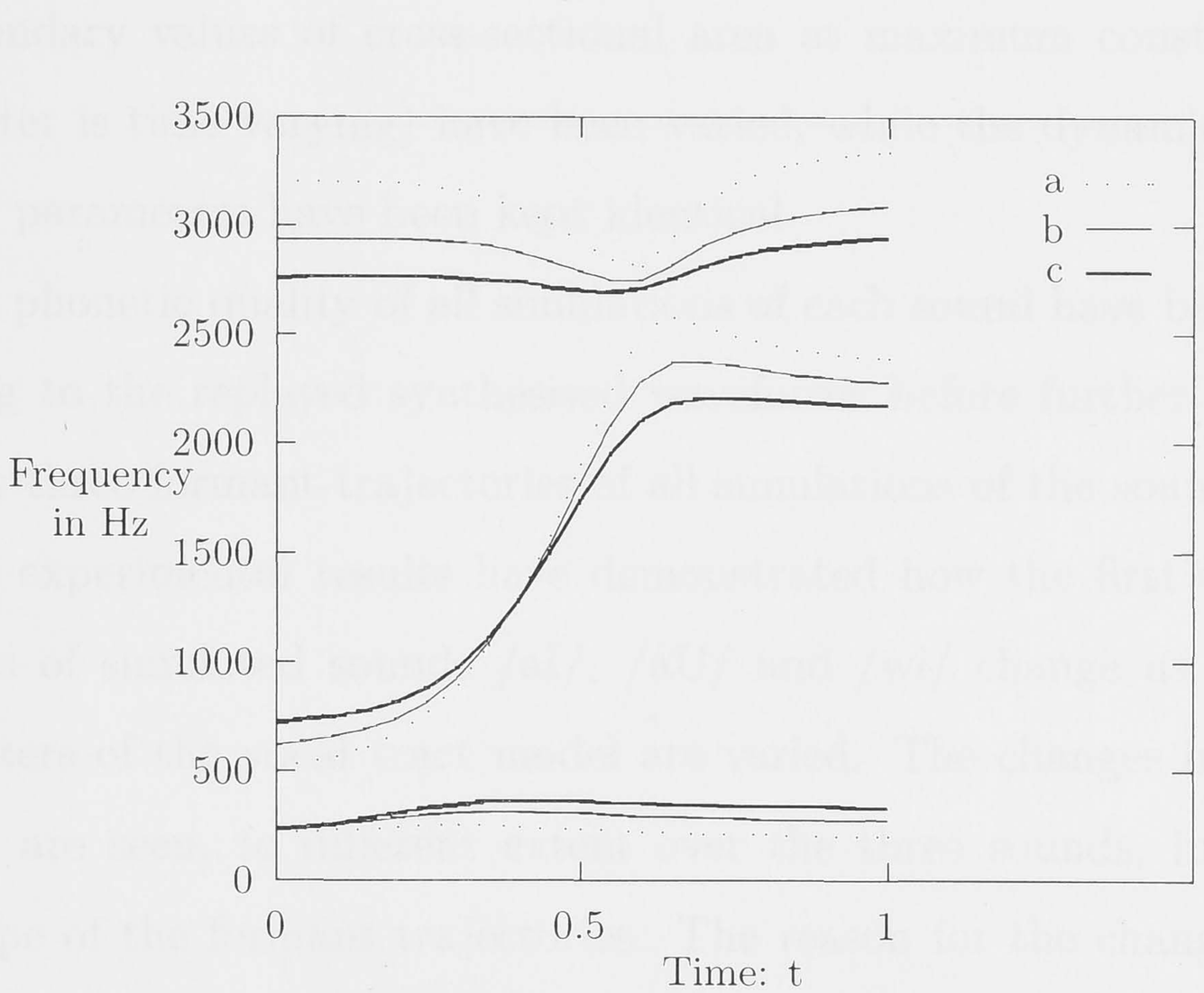


Figure 4-22: The first three formant trajectories for /wi/ with variation of A_c, a: $A_c = .289$, b: $A_c = .65$, c: $A_c = 1.16$.

4.8 Summary

This chapter reports the tests on the hypothesis that differences in size and shape of the vocal tract also influence the dynamics of the formant trajectories in the speech signal.

Three dynamic phonetic segments: /aI/ as in "high", /aU/ as in "how" and /wi/ as in "we", have been simulated. In the simulation, an acoustic tube model of the vocal tract, transmission line theory, digital signal processing techniques were applied. For each sound, two particular parameters of the model were chosen as time varying and following a sigmoidal or piecewise sigmoidal function.

The values of model parameters for the benchmark simulation of each sound were chosen generally at the level for a male speaker. Based on the benchmark parameter values, for each sound, the size and shape of the vocal tract has been perturbed. That is, 1) the length of the pharyngeal portion, 2) the maximum cross-sectional area, 3) cross-sectional area at maximum constriction (or boundary values of cross-sectional area at maximum constriction if this parameter is time varying) have been varied, while the dynamics of the time varying parameters have been kept identical.

The phonetic quality of all simulations of each sound have been verified by listening to the replayed synthesised waveforms before further assessment of the first three formant trajectories of all simulations of the sounds.

The experimental results have demonstrated how the first three formant patterns of simulated sounds /aI/, /aU/ and /wi/ change as certain static parameters of the vocal tract model are varied. The changes in the formant pattern are seen, to different extent over the three sounds, in the position and slope of the formant trajectories. The reason for the changes is because the effect of the static condition of the vocal tract on the first three formant values at different positions on the trajectory of the articulator movement for the sound is not uniform, reflecting another perspective of the nonlinear

relationships between articulatory parameters and acoustic parameters. In the results, trajectory features which may intuitively be associated with the dynamics of articulation are clearly seen to be consequences of the interaction of articulatory dynamics and the static condition of the articulation.

Further discussions and conclusions are given in Chapter 6.

Chapter 5

On the Separation of Speech Signal Variances from Two Sources

5.1 Introduction

In speech science research, the study on static sounds, e.g. English monophthongal vowels, has a history longer than that on dynamic sounds (i.e. the sounds, the acoustic signals of which vary with time). For those static sounds, their spectral properties are supposed to be stationary (not change with time). If we look at acoustic variance of a sound caused by speaker variation of a certain group people, we would lump the values of the spectral parameters of the repeated tokens by those people together and calculate the statistics such as mean and variance. If we look at the acoustic variance caused by speaker variation occurring in dynamic sounds, it is obviously incorrect to analyse it in the same way as for static sounds. In this case, the acoustic variance caused by speaker variation (this variance is also called speaker variance in the thesis) should be separately represented from the acoustic variance caused by phonetic variation (this variance is also called acoustic dynamics/phonetic

variance). The question here is how do we separate these two variances to a certain degree in order to study each variance individually, such as

- to compare the differences in acoustic dynamics realisations of a phonetic variation and intra-speaker variance between speakers, or
- to compare the differences in acoustic dynamics realisations of a phonetic variation and intra-speaker variance between different physiological or emotional states of individual speakers, or
- to compare the differences in acoustic dynamics realisations of a phonetic variation and speaker variance between different sounds.

Hidden Markov modeling is a mathematical statistical method to model a non-stationary process. It has now been widely adopted in speech/speaker recognition systems to separately model the acoustic variance due to phonetic variation and the variance due to variation from other sources such as speaker variation (e.g. Poritz 1988, Rabiner 1989, Tishby 1991). In a hidden Markov model, a time varying speech signal is globally modeled by a sequence of states representing temporal change of the acoustic properties of the speech event. Within each state, the associated local characteristics are modeled by a probability distribution. In principle, this method suits the purpose of separation of the acoustic variances from phonetic variation and other speaker related variation.

However, in the speech technology systems, the number of states in each model is usually selected intuitively before modeling and to be the same over models of different speech events in the system. It can be expected that if the number of states in a model is small, the variance in an individual state reflects not only inter- and/or intra-speaker variance, but to a certain extent phonetic variance as well. The detail of acoustic dynamics due to phonetic variation will increase and the influence of phonetic variance on the measure of speaker variance will decrease as the number of states in a model increases.

Theoretically, the number of states can be chosen to be as many as that with which if one more state is added there will be at least a pair of adjacent states of the model having the same distribution (or they have no significant difference). Different physical events modeled (e.g. different English words, different speakers) may have different acoustic dynamics and inter- and/or intra-speaker variances. When the number of states is chosen to be the same over all models, the resolution of the acoustic dynamics and the influence of the acoustic dynamics on speaker variance in individual states may be different from speech sound to speech sound and/or from speaker to speaker. Thus in this situation, any study on acoustic dynamics and speaker variance is difficult to conduct.

In speech technology research, three studies on the extensively used "cepstral" space for distance measures of template based speech/speaker recognition have calculated cepstral variances based on (a) a wide range of sounds from several speakers (Juang *et al.* 1987), (b) a single word from a large range of speakers (Tohkura 1987), and (c) some isolated and connected words repeated several times by individual speakers (Soong *et al.* 1988).

In their study, Juang *et al.* (1987) have tried to demonstrate the components in the cepstral variance that are primarily artifacts of the LPC analysis procedure and are simply not useful for the speech recognition. They measured the variance of each of LPC cepstral coefficients for the speech signals including a wide range of sounds from 7 speakers and labeled as the variances for "mixed data". They also measured the variance of each of LPC cepstrals of a synthesised vowel-like sound from an all-pole filter excited by a zero mean Gaussian distributed signal and labeled as the variance for "simulated data". They compared the differences of these two kinds of variance by looking at the ratio of that of each cepstral coefficient for the "simulated data" to the corresponding one for the "mixed data". They showed that the ratio for the high order cepstrals are relatively greater than that for the low order cepstrals. They argued that, this increase with cepstral index indicates the diminishing

discriminating power of the higher order cepstrals and the variance of those cepstrals are inherent artifacts of the analysis procedure and hence they are less desirable in spectral similarity comparisons than lower order cepstrals. They proposed a filtering window used to suppress the undesirable cepstral coefficients for an automatic speech recognition system.

Here it should be noted that their measured variance for the "mixed data" includes inter-speaker variance, intra-speaker variance, and phonetic variance. If their measured variance for the "simulated data" is regarded to represent the inter- /intra-speaker variance that is associated with repetition of the same vowel sound in normal physiological and emotional states, their observed difference may be due to the phonetic variance embedded in the "mixed data".

In the speaker recognition research, Soong *et al.* (1988) have applied the Mahalanobis distance measure defined in the Gaussian distribution of mathematical statistics to a VQ-based speaker recognition system. They argued for using this measure to "*equalise the contributions of individual cepstral coefficients to the performance of VQ based recognition*". They use all training data from a speaker to calculate the covariance matrix for that speaker and call it *pooled intra-speaker covariance*. This *pooled intra-speaker covariance* actually includes phonetic variance and intra-speaker variance.

Their use of the Mahalanobis distance could be explained as trying to lift the modeling constraint of an "identity" covariance matrix for each codeword (see Section 3.2: Comparison of VQ & GM Models). If so, a covariance matrix should be estimated for each codeword based on the data samples being allocated to that codeword. The data within a codeword take a small portion of the region occupied by all training data in the cepstral space. If the speech data cover a certain wide range of phonetic sounds and the speech signals are repeatedly produced by a speaker under normal physiological and emotional states (it is often true when training a speaker recognition system), the covariance of a codeword could be dominated more strongly than *pooled intra-speaker covariance* by intra-speaker repetition variance. If there were

substantial differences between codeword's covariance and *pooled intra-speaker covariance*, we should not replace the covariance matrix of a codeword with the *pooled intra-speaker covariance.*

This chapter reports the methods developed, based on hidden Markov modeling techniques to separately model monotonic acoustic dynamics (the acoustics changes monotonically with time) due to phonetic variation and the acoustic speaker variance occurring within that dynamics. This separate modeling was designed to meet two criteria: (1) within each model influence of the acoustic dynamics on the variance of individual state should be reasonably uniform across all states, and (2) the average degree of influence of the acoustic dynamics on the variance of individual state should be similar across different modeled events (see next section for details of the methods). The methods have been applied to the monosyllabic words database described in Chapter 2 to test the methods. The vocalic part of "we", "high" and "how" per speaker (i.e. one model per sound/speaker pair) were modeled in LPC cepstral space. Hence, the acoustic dynamics due to phonetic variation and intra-speaker repetition variance are the main features to be modeled. With the methods to extract the variance caused by intra-speaker repetition variation, it has been able to investigate the differences between the intra-speaker repetition variance and the variance, which is the mixed of variances from the two sources (i.e. the phonetic variation and the intra-speaker repetition variation), in each of the sound/speaker pairs. Section 5.3 reports the experimental results. A summary is given in Section 5.4.

5.2 Methods

This section describes the procedure to separately model the acoustic dynamics and acoustic intra-speaker variance of each sound/speaker pair under the framework of hidden Markov modeling. The procedure comprises three basic parts: (1) a model and its structure; (2) a criterion for estimation; and (3)

an algorithm to implement the estimation. These parts are discussed in Subsection 5.2.1, 5.2.2, and 5.2.3 respectively. The novel methods for forming the initial model and for selection of the number of states of a model, which were developed to meet the criteria of the modeling in this study, are described in Sub- section 5.2.4 and Sub-section 5.2.5.

5.2.1 The Model and Its Structure

A hidden Markov model can be summarised by its parameters $\lambda = (A, a_0, B)$, where

- $A = \{a_{ij}, i, j = 1, ...N\}$ is the state transition probability matrix;
- a_0 is the initial state probability distribution;
- $B = \{b_i, i = 1, ...N\}$ are output probability distributions for each state i;
- N is the number of states in a model.

The simple left-to-right structure with no skipped states (Figure 5-1) has the property that as time increases the state index either increases sequentially or stays the same. If it is assumed that all observation sequences of a given sound/speaker pair have the same state sequence, then this model structure can be used to model the given sound/speaker pair. The state sequence represents the acoustic dynamics due to phonetic variation, and each state describes, by its output distribution, the corresponding acoustic properties of one phonetic state of the utterance.

The adopted model structure leads to $a_{ij} = 0(i > j; i < j - 1)$ in A. The self-transition probability a_{ii} conveys information about the duration of state i although in this architecture it does not model duration accurately (see Russell *et al.* 1985, Levinson 1986). If we make all non-zero values in the transition probability matrix equal, the duration information is not included

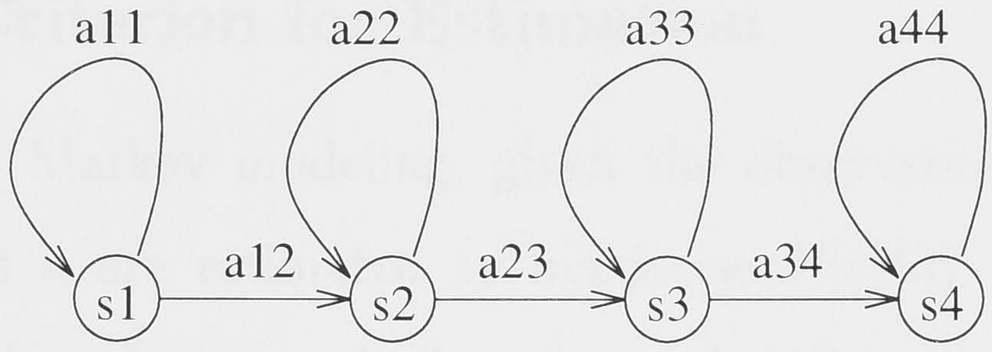


Figure 5-1: Structure of a 4-state left-to-right hidden Markov model

in the model. When the Viterbi algorithm is used during modeling (see Subsection 5.2.3) however, the allocation of each observation into either current state or the next state will only depend on which state the data itself matches closer. Consequently, the estimated distribution of individual state represents the properties of the local data. Due to the mathematical constraint that the sum of transition probabilities from any state should be unity, all values are set to 0.5 so that the constraint is obeyed.

Taken together, the above considerations result in a state transition probability matrix of the following form, with $N = 4$ for example:

$$A = \begin{pmatrix} 0.5 & 0.5 & 0.0 & 0.0 \\ 0.0 & 0.5 & 0.5 & 0.0 \\ 0.0 & 0.0 & 0.5 & 0.5 \\ 0.0 & 0.0 & 0.0 & 1.0 \end{pmatrix}$$

An initial state probability distribution a_0 of 1 for state one and 0 for other states restricts each observation sequence to starting with state one.

The output distribution for each state b_i is assumed to follow a multivariate Gaussian distribution. Hence the variance within a state, in the form of the covariance matrix of this multivariate Gaussian, could be regarded as representing intra-speaker repetition variance for the corresponding phonetic state.

The number of states N is not chosen the same for all models. See Subsection 5.2.5 for details of the method for choice of N.

5.2.2 The Criterion for Estimation

In general hidden Markov modeling, given the observation sequence O, the model parameters λ are estimated to maximise $P(O|\lambda) = \sum_{s \in S} P(O, s|\lambda)$. This means that λ is chosen such that the probability of O with all possible state-labeled observation sequences S on λ is maximised.

An alternative criterion, which better suits the application for this study, is to maximise the joint likelihood $P(O, s_m|\lambda) = max_{s \in S} P(O, s|\lambda)$. This means that λ is chosen such that the likelihood of O with s_m on λ is maximised, where s_m is the most likely among all possible state-labeled sequences S of O.

Given all observation sequences of a sound/speaker pair, we are interested in the model on which the probability of each observation sequence with s_m is maximum. Then with s_m, all data labeled as the same state can be regarded to represent the acoustic properties of a phonetic state in that sound/speaker pair. Therefore, it was decided to select maximum joint likelihood $maxP(O, s_m\lambda)$ as the criterion for estimation of model parameters λ.

5.2.3 The Algorithm

The development of a particular algorithm for estimation of model parameters arises from the interaction of selection of the estimation criterion and the model type.

The Baum-Welch algorithm (e.g. Rabiner 1989) is appropriate for estimation of hidden Markov model parameters λ in the sense of maximum likelihood $maxP(O|\lambda)$. In the study, because the criterion function is $P(O, s_m|\lambda)$, the segmental K-means algorithm (Rabiner *et al.* 1985) was used to estimate λ by maximising $P(O, s_m|\lambda)$.

Since the initial probability distribution a_0 and the state transition probability matrix A have been fixed, the segmental K-means algorithm is used only to estimate parameters of the output distribution.

The segmental K-means algorithm 1) uses the Viterbi algorithm (Forney

1973) to segment each observation sequence of a sound/speaker pair into states; 2) makes a maximum likelihood estimation of the mean vector and covariance matrix for each state based on the observations segmented, from all observation sequences, into that state. This procedure is repeated until the criterion function converges.

5.2.4 Initial Model

As the segmental K-means algorithm is itself an iterative estimation procedure, the initial model is significant for the convergence to the final estimated model. The initial model is usually (e.g. Rabiner, 1989; Hidden Markov Model Toolkit V1.5, 1993) created from a uniform segmentation in time of each observation sequence into states, as for example shown in Figure 5-2 (b). If an observation sequence is a nonlinear function of time, a uniform segmentation in time will lead to a fact that some states are more strongly influenced by the acoustic dynamics than other states (see Figure 5-2 (b)).

In order to improve this situation, a simple method was developed for forming the initial model. The method segments each observation sequence into states according to its dynamic properties. With this method, the segments of each observation sequence would have identical dynamic range. It involves three steps:

- Step 1: the sum of Euclidean distances between each pair of observations (frames) adjacent in time from an observation sequence is computed;
- Step 2: given the number of states n, a threshold is chosen by dividing the sum obtained in Step 1 by n;
- Step 3: starting with the first observation of the sequence, Euclidean distances for each successive pair of adjacent observations are accumulated until the sum reaches the chosen threshold. All observations before this point (when the sum of Euclidean distances exceeds the threshold), are

allocated to state number one. This procedure is restarted and continued until all observations have been examined and allocated to a corresponding state.

In reality, the acoustic observation sequence of the speech is usually a nonlinear function of time. Initial models derived using the method (Figure 5-2 (a) gives an example) should help to obtain the models in which influence of the acoustic dynamics on the variance of individual state across all states is more uniform than the models obtained via uniform segmentation in time.

5.2.5 Choice of the Number of States N

In order to meet the second criterion of the modeling, a novel method for the selection of the number of states (N) was developed. In the method, the relative distance between any two adjacent states i and $i+1, i = 1, ..., n-1$ in each trial model is measured, which is defined as

$$D_i^2(n) = (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_{i+1})' \mathbf{S}_p^{-1} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_{i+1})$$

where $\bar{\mathbf{x}}_i$ and $\bar{\mathbf{x}}_{i+1}$ are the mean vectors of state i and state $i+1$ respectively. $\mathbf{S}_p$ is the averaged covariance of $\mathbf{S}_i$ and $\mathbf{S}_{i+1}$, which are the covariances of state i and state $i+1$ respectively.

$D_i^2(n)$ reflects modeling resolution of the acoustic dynamics and the degree of separation between the acoustic dynamics and intra-speaker repetition variance. For the speech signals, the sequence of acoustic parameters of which shows its monotonic relation to time, the model with greater number of states will have less value of $D_i^2(n)$. Given trial models with the same n, greater acoustic dynamics (more change of the acoustic with time) of a sound/speaker pair will lead to larger distance between mean vectors $\bar{\mathbf{x}}_i$ and $\bar{\mathbf{x}}_{i+1}$ of a pair of any two adjacent states. Not only is acoustic dynamics modeled at a relatively lower resolution, but also the variance in individual states is more influenced by the acoustic dynamics. Large value of $D_i^2(n)$ in this case indicates that a

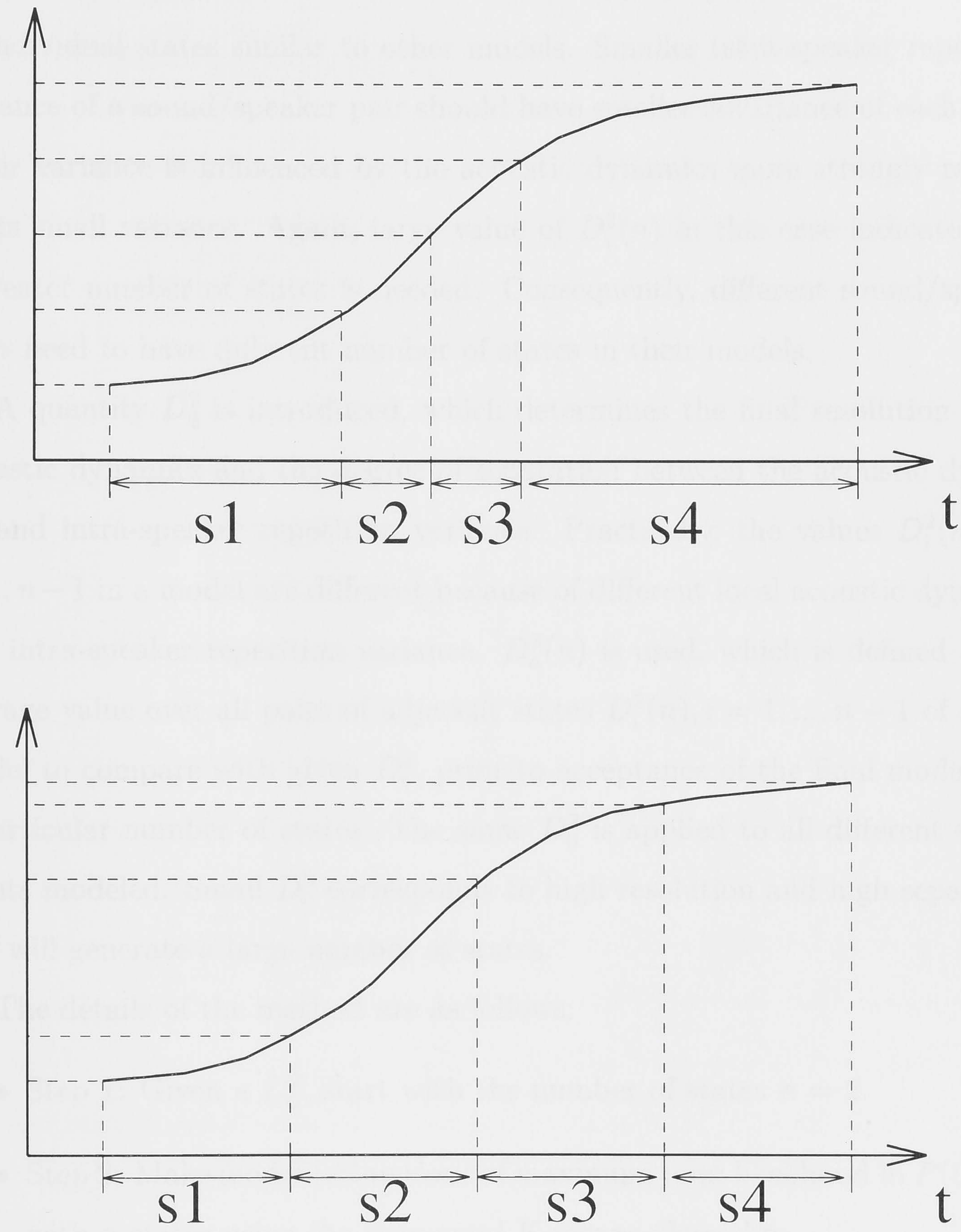


Figure 5-2: Illustration of segmentation of an observation sequence as a function of time. The top (a) is uniform segmentation in the cepstral domain. The bottom (b) is uniform segmentation in time domain. Horizontal axis represents time and the vertical can represent one dimension of cepstral coefficients.

greater number of states is needed in order to make the resolution of modeled acoustic dynamics and the influence of acoustic dynamics on the variance of individual states similar to other models. Smaller intra-speaker repetition variance of a sound/speaker pair should have smaller covariance of each state. Their variance is influenced by the acoustic dynamics more strongly relative to its small variance. Again, large value of $D_i^2(n)$ in this case indicates that a greater number of states is needed. Consequently, different sound/speaker pairs need to have different number of states in their models.

A quantity D_0^2 is introduced, which determines the final resolution of the acoustic dynamics and the degree of separation between the acoustic dynamics and intra-speaker repetition variance. Practically, the values $D_i^2(n), i = 1, ..., n-1$ in a model are different because of different local acoustic dynamics and intra-speaker repetition variance. $D_a^2(n)$ is used, which is defined as the average value over all pairs of adjacent states $D_i^2(n), i = 1, ..., n-1$ of a trial model to compare with given D_0^2, prior to acceptance of the final model with a particular number of states. The same D_0^2 is applied to all different speech events modeled. Small D_0^2 corresponds to high resolution and high separation and will generate a large number of states.

The details of the method are as follows:

- Step 1: Given a D_0^2, start with the number of states $n = 2$.
- Step 2: Make model estimations of maximum joint likelihood in $P(O, s|\lambda)$ with n states using the segmental K-means algorithm.
- Step 3: Compute relative distance between every two adjacent states $D_i^2 n, i = 1, ..., n-1$ and the average value over them, $D_a^2(n)$.
- Step 4: Calculate $D^2(n) = |D_a^2(n) - D_0^2(n)|$.
- Step 5: Compare $D^2(n)$ with $D^2(n-1)$. If $D^2(n) > D^2(n-1)$ (it means $D_a^2(n-1)$ is the closest to D_0^2), go to Step 7.
- Step 6: Increase n by 1, go to Step 2.

- Step 7: Stop. The final number of states for the model is $N = n - 1$.

5.3 Experimental Investigations

The methods have been applied to model the vocalic part of three diphthongal sounds "we", "how", and "high" per speaker in the monosyllabic word database. A total of 33 left-to-right hidden Markov models with $D_0^2 = 4$ were obtained. $D_0^2 = 4$ indicates that in each model on average the distance between any two adjacent states is equal to the sum of their standard deviations. Table 5.1 lists only the number of states for each of the 33 models. Table 5.2 lists the number of states for each of the 33 models trained with $D_0^2 = 8$ (double the D_0^2 value used in obtaining the values on Table 5.1), and each value is less than the corresponding one in Table 5.1. The result is consistent with the expectation that the number of states required for a model decreases as D_0^2 increases. It means that the resolution of phonetic variance modeled is lower (small N) and acoustic dynamics more strongly influence measures of intra-speaker repetition variance internal to each of the N states.

Given a D_0^2 (e.g. either $D_0^2 = 4$ or $D_0^2 = 8$), it is observed that for almost all speakers, the number of states of the model for sound /wi/ is larger than that of the other two sounds. This indicates that sound /wi/ has greater dynamics in its spectrum provided the intra-speaker repetition variance for those sounds are the same. Recalling the three simulated sounds in the study on the sources of inter- and intra- speaker variability in the acoustic dynamics of speech (Chapter 4), it is known that the sound /wi/ has greater change in its first three formants than the sounds /aI/ and /aU/. Because the first three formants dominate the corresponding spectrum, it can be expected that sound /wi/ has greater change in its spectrum than the other two sounds. This is consistent with the modeling results.

With same D_0^2, for each particular sound, variation of the number of states across speakers, together with the mean vectors and covariance matrices, sug-

gests their differing acoustic realisation of the sound and/or intra-speaker repetition variance.

speaker	number of states		
	we	*how*	*high*
AH	17	11	14
BM	16	14	12
DD	18	15	16
GC	15	12	14
IM	20	13	16
JW	16	12	13
KR	21	18	19
NF	18	11	13
SB	16	13	16
TB	21	15	16
WB	15	12	15

Table 5.1: The number of states for each model of 33 sound/speaker pairs with $D_0^2 = 4$.

In order to investigate the differences between the intra-speaker repetition variance and the variance which is a combination of the variances from the two sources (i.e. the phonetic variation and the intra-speaker repetition variation), the covariance has been computed over all data samples for each sound/speaker pair. The sum of the diagonal elements for each of 33 pairs is listed in Table 5.3. In fact, each value of Table 5.3 is a combination of phonetic variance and intra-speaker repetition variance, which here is labeled as the global variance. The averaged covariance was also calculated over all states in each of the 33 models with $D_0^2 = 4$. The sum of the diagonal elements of the averaged covariance matrix, representing the average variance per state of the model is listed in

speaker	number of states		
	we	*how*	*high*
AH	10	6	6
BM	12	6	7
DD	12	8	8
GC	10	6	7
IM	13	7	8
JW	12	7	6
KR	12	10	10
NF	13	6	8
SB	10	6	7
TB	12	6	8
WB	9	7	8

Table 5.2: The number of states for each model of 33 sound/speaker pairs with $D_0^2 = 8$.

Table 5.4. It can be seen that the measure in Table 5.4 is in nearly every case less than half of the corresponding global variance in Table 5.3, and in quite a number of cases is of the order of a quarter of that in Table 5.3. Each measure in Table 5.4 is regarded as the intra- speaker repetition variance, because the phonetic variance components in the measure has been reduced to a certain degree when compared with the global variance. It is the phonetic variance that causes the global variance to be greater than the intra-speaker repetition variance.

speaker	variance		
	we	*how*	*high*
AH	.657	.392	.362
BM	.871	.419	.458
DD	.770	.370	.481
GC	.631	.437	.366
IM	.747	.392	.597
JW	.729	.405	.500
KR	.818	.507	.760
NF	1.194	.487	.673
SB	.645	.397	.436
TB	.699	.369	.605
WB	.633	.332	.343

Table 5.3: The global variance for each of 33 sound/speaker pairs, which is the sum of the diagonal elements of the covariance matrix derived from all observations.

A detail comparison between the global variance and the intra-speaker repetition variance in the cepstral space for each sound/speaker pair has been further conducted. The global variance and the intra-speaker repetition variance per cepstral dimension for the three sounds by speaker AH as one example are

speaker	average variance/state		
	we	*how*	*high*
AH	.204	.198	.186
BM	.196	.189	.180
DD	.196	.153	.183
GC	.160	.161	.188
IM	.181	.154	.150
JW	.196	.179	.193
KR	.179	.137	.180
NF	.213	.232	.228
SB	.179	.179	.206
TB	.211	.167	.196
WB	.148	.160	.145

Table 5.4: The intra-speaker repetition variance of each of the 33 models with $D_0^2 = 4$. It is the sum of the diagonal elements of the covariance matrix averaged across all states in a model.

shown in Figure 5-3. The figures for the variances of the rest of sound/speaker pairs in this study are included in Appendix B. The upper plot in each of figures is the global variance. For each sound, similarity can be seen in the curvature of the global variance against cepstral index between the speakers. For instance, there is a peak apparently at index 5 of the cepstrals for sound /aI/ for all speakers. For each speaker, dissimilarity is also seen in the curvature of the global variance against cepstral index between the sounds. The observed similarity and dissimilarity reflect influence of the phonetic variance on the global variance. Both variance values generally decrease with increase of the cepstral index. However in most cases, the difference between the value in the upper plot and the value in the lower plot decreases (or the ratio of the value in the lower plot to the value in the upper plot increases) with increase of the cepstral index. The greater the value in the upper plot than that in the lower plot, the more strongly it is dominated by the phonetic variance in that dimension of cepstrals.

5.4 Summary

The methods have been developed, based on hidden Markov modeling using an iteratively designed architecture, to separately model the monotonic acoustic dynamics and speaker variance. A method of uniform segmentation of each observation sequence into states in cepstral domain when forming the initial model was developed to meet first criterion of this separate modeling: the influence of the acoustic dynamics on the variance of individual state (representing speaker variance) is reasonably uniform across states in a model. The relative distance between any two adjacent states of a hidden Markov model was introduced and the same criterion for the averaged relative distance across states in a model was applied to meet second criterion of the separate modeling: the influence of the acoustic dynamics on the speaker variance is similar over different speech events modeled. With the methods developed, if the speaker

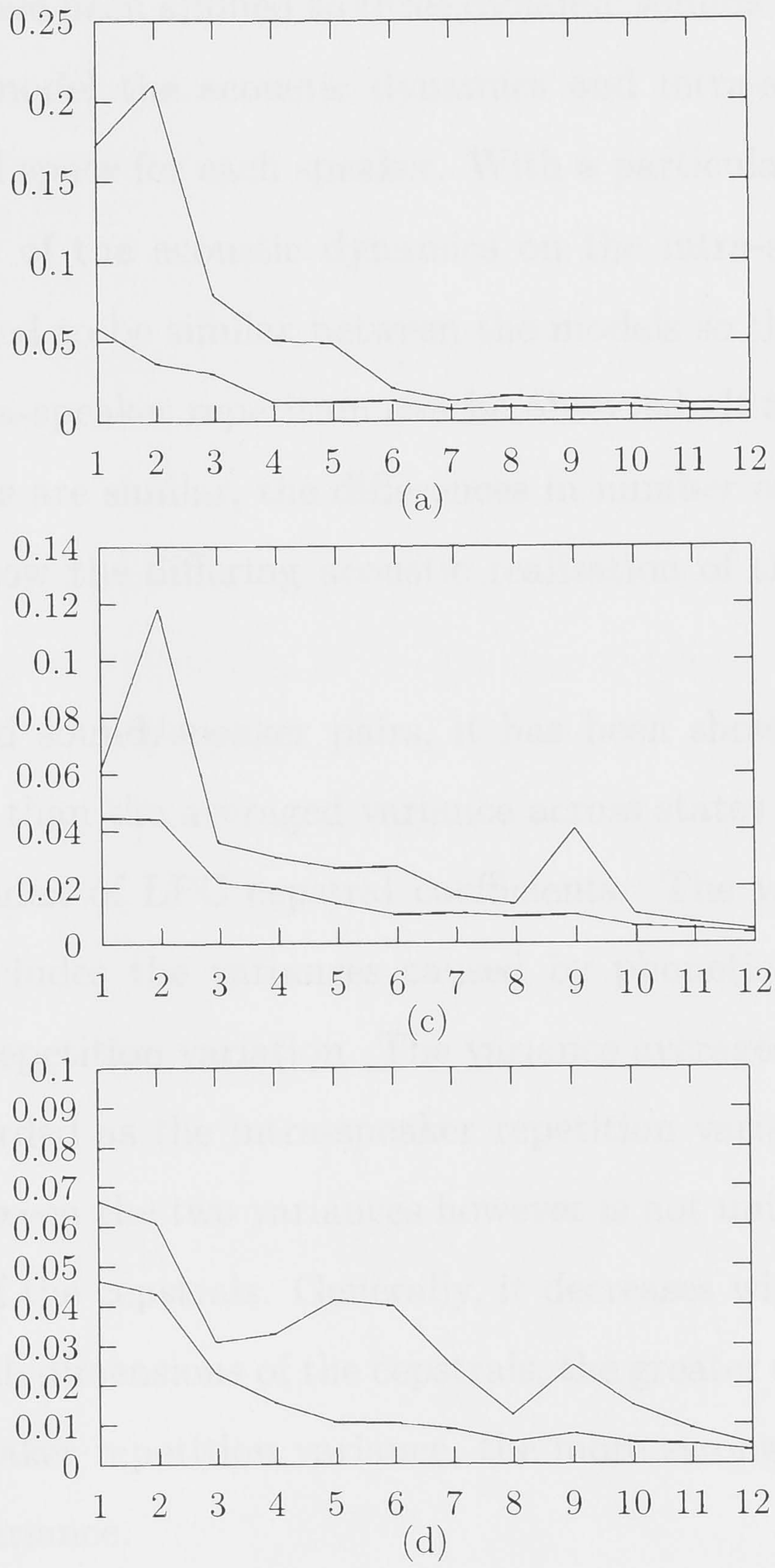


Figure 5-3: The global variance and intra-speaker repetition variance against each of 12 LPC cepstral coefficients; (a): "we", (c): "how", (d): "high" by speaker AH. The upper curve is the global variance. The index of the cepstral coefficients goes from 1 to 12.

variances are the same between different speech events modeled, the resolution of the acoustic dynamics will be modeled reasonably similar between the speech events.

The methods have been applied to three dynamic sounds by 11 male speakers to separately model the acoustic dynamics and intra-speaker repetition variance in cepstral space for each speaker. With a particular value of D_0^2, the degree of influence of the acoustic dynamics on the intra-speaker repetition variance is controlled to be similar between the models so that differences between them in intra-speaker repetition can be observed. If their intra-speaker repetition variances are similar, the differences in number of states and mean vector sequence show the differing acoustic realisation of the phonetic variation.

For the selected sound/speaker pairs, it has been shown that the global variance is greater than the averaged variance across states in the model with $D_0^2 = 4$ in each index of LPC cepstral coefficients. The variance labeled as global variance includes the variances caused by phonetic variation and by the intra-speaker repetition variation. The variance averaged across the states of a model is regarded as the intra-speaker repetition variance in this study. The difference between the two variances however is not uniformly distributed across the index of the cepstrals. Generally, it decreases with increase of cepstral index. Over all dimensions of the cepstrals, the greater the global variance than the intra-speaker repetition variance, the more strongly it is dominated by the phonetic variance.

Further discussions and conclusions are given in Chapter 6.

Chapter 6

Concluding Discussions

The previous chapters have reported in detail the thesis studies to tackle some scientific and engineering problems in the automatic speaker recognition. This chapter first discusses in a broad way the findings of the studies reported in the past three chapters. Then it generalises from the thesis work on the way that the speaker recognition research should go. Finally, it summarises the contributions of this thesis work to automatic speaker recognition area and, in general, to speech science and engineering area.

6.1 On Speaker Modeling for the Speaker Recognition

6.1.1 Choice of Model Form

Selection of model form in speaker modeling for the speaker recognition has been shown to be an important issue. VQ model and Gaussian mixtures (GM) model are most popularly selected ones, which are usually regarded as different forms. Rather than following the usual paradigm of assessing their speaker recognition performance, the study has: (i) explicated mathematical relationships between the two models, showing that VQ model is equivalent to GM model with three imposed constraints: 1) non-overlapping between

components, 2) equal mixture weights, 3) identity of covariance matrix of each component, and (ii) experimentally demonstrated that the three constraints on the GM are not applicable to real speech data, suggesting that GM model should provide more effective means than the VQ to accurately model a speaker for the speaker recognition.

The use of VQ model and GM model in the speaker modeling is not because people know that the speech signals from any speaker follow those models, but because those model forms can be flexibly used to approximate any arbitrary distributions. One example is that many studies have experimentally investigated the system performance with different number of the codewords or the mixtures to seek for the optimal number suggested for the system. Because GM model releases the three constrictions on the speech data, it has more freedom to closely fit to the real distribution. The trade-off is that in representing the real distribution, GM model has more parameters to be estimated. With the Bayesian decision rules, the correct representation of real distributions is a sufficient condition to achieve optimal performance of the recognition tasks. The study suggests that we should study characteristics of the speech signals from a speaker and find a model form with fewer parameters but closer fit to the real distribution.

6.1.2 The Modeling Approach

Model estimation method has been identified as another important issue in model based speaker recognition. This is true particularly when correctness of the model form selected for the recognition system is challenged. The study has reviewed two popular methods proposed to replace the maximum likelihood estimation method (the MLE) when the assumption on correctness of model form is invalid: 1) the maximum mutual information estimation method (the MMIE) in which a new statistic is defined to be maximised, 2) the corrective training method which intuitively follows the error correcting training procedure for linear classifiers.

The analysis of the speaker identification experiments in the study suggests that, if the mismatch between the real distribution of a speaker's data and estimated model by the MLE causes recognition errors, those errors may occur in the regions, in which the differences between a speaker's own distribution and other speakers' distributions are relatively small and ranking relations of these distributions are not correctly represented. Correct ranking order of the probabilities between speakers' models is essential for the optimal performance of Bayesian rule based recognition tasks, no matter how accurate the models fit to the real distributions.

An innovative estimation method has been developed to improve correctness of the ranking order between the models estimated by the MLE, hence to improve the identification performance. This improvement has been shown with the GM model form and a set of five male speakers. It has also been shown that models estimated by the innovative estimation method are a poorer fit to the real distributions than the models estimated by the MLE in maximum likelihood sense.

It has been argued that the models further re-estimated by the MMIE will remain unchanged if they are initially estimated by the MLE and at least one model has its single maximum value (this is more likely a case with the real speech data). It has been experimentally demonstrated that models with greater discrimination in the sense of maximum mutual information as defined in the MMIE do not necessarily lead to better identification performance.

The reason why the corrective training method is so effective as reported by its developers (Bahl *et al.* 1988) is because the method partly works to improve correctness of the ranking order of probabilities between the models. The difference between the corrective training method and the innovative estimation method lies in that, the former intuitively follows the error correcting training procedure for linear classifiers and the latter is purposely designed to correct the ranking order in the regions caused by the use of incorrect model form. Because of this difference, there are some technical differences between

the two methods as described in Sub-section 3.3.5.

Ideally, it would be highly encouraged to study characteristics of the speech signals from a speaker and to seek for a model form which is able to represent the real distribution more accurately. The closer the models are to the real distributions, the more likely the ranking order between the real distributions could be correctly represented. Practically, a model estimation strategy can be suggested from the study for the situation when we are not confident with correctness of the selected model form. First, the MLE method is used to obtain the models which match the real distributions as closely as possible. Then, the innovative estimation method can be used to adjust model parameters in order to improve correctness of the ranking order. The adjustment will reduce the degree of model accuracy to the real distributions. But if the recognition performance with the training data were to increase, it would also indicate that correctness of the ranking order is improved.

In the speaker verification, a fixed threshold is usually set on a speaker's own model to decide acceptance or rejection of the claim for him. The absolute value affects the type I errors and type II errors in opposite way. The greater the value is, the less the type I error rate but greater the type II error rate. In order to minimise both type I and II errors, it is suggested to introduce the imposter's model for each speaker. The model may be estimated from the speech data by other speakers assumed as imposters. Then verification decision could be made by equivalently making identification between the true speaker and the his impostor based on the Bayes rules. If the probability of test data on the speaker's model is greater than that on the impostor's model, the decision will be acceptance, otherwise the rejection. The type I and II errors would reach their minimum if the speaker's model and the imposter's model represented their real distributions. Consequently, the innovative modeling method discussed in Chapter 3 could be adopted to reduce both type I and II errors due to incorrectness of model form used.

6.2 On the Sources of Inter- & Intra- Speaker Variability in the Acoustic Dynamics of Speech

The purpose of the study reported in Chapter 4 is to test the hypothesis that differences in size and shape of the vocal tract also influence the dynamics of the formant trajectories in the speech signal. For this purpose, three dynamic phonetic segments (/aI/ as in "high", /aU/ as in "how" and /wi/ as in "we") were simulated using an acoustic lossless tube model. In simulation of each sound, the values of some static parameters of the model were perturbed, but the dynamics of the time varying parameters were kept unchanged. The perturbed parameters are length of the pharyngeal portion (hence the vocal tract length L), the maximum cross sectional area of the vocal tract(A_m) and the cross sectional area of maximum constriction point (A_c). The study has demonstrated how the first three formant trajectory patterns of simulated sounds change as those parameters are varied. The changes in the formant pattern are seen, to different extent over the three sounds, in the position, shape and slope of the formant trajectories. The results support the hypothesis.

Tosi *et al.* (1972) recognised the importance of slopes of formants during liquids, glides and diphthongs in the speaker identification from visual examination of spectrograms. Stevens (1971) regarded the sources of variability in the slope of the formants for those sounds as "*may be the result of learned articulatory habits rather than differences in anatomical or physiological characteristics of the articulatory structure*" (note he relates articulatory dynamics to *learned articulatory habits*). In studying the acoustic features for the speaker identification, Sambur (1975) claims, "*The F_2 slope in /aI/ was also quite variable among speakers and demonstrated excellent identification potential*", which is one of "*measurements related to the dynamic properties of the talker's voice patterns that reflect his learned behavior of speaking*". Certainly,

different articulatory dynamics causes differences in the corresponding acoustic dynamics. However the experiments in simulating diphthongal sounds of the study have shown that variability in the slope of the first three formant trajectories can be the results of different size and shape of the vocal tract. As one example, when the cross-sectional area contrast (i.e. $\frac{A_m}{A_c}$) of the vocal tract is increased in producing /aI/, the slope of $F_2(t)$ increases.

There are differences between the changes in patterns of the first three formant trajectories due to the differences in the anatomical structure and the changes due to differences in the articulatory dynamics. The slope of the formant trajectories will increase if the articulators move faster. However the range of formant values during the trajectories and the formant trajectory patterns will be kept the same if there is no change in the articulatory static aspects. When the size or shape of the vocal tract varies, both the pattern and range of the formant trajectories would have changes.

Based on their 3-parameter vocal tract models, Stevens & House (1955) and Fant (1960) have studied the relations between the articulatory parameters and the acoustic parameters, which are intrinsically determined by the vocal tract geometry. One example is the relation between the maximum constriction position and formants when constriction degree and lip rounding are fixed. Some relations have shown nonlinearity, and have been suggested to play a significant role in shaping distinctive features of speech sound system by Stevens (1972, 1989). Also based on a 3-parameter model, this study has revealed another perspective of the nonlinear relationship between the articulatory parameters and the acoustic parameters of the vocal tract. It has been demonstrated that the size and shape of the vocal tract affect the values of formants non-uniformly along the particular regions the articulators pass through when producing sounds. If the articulators move through those regions with time, this effect will be manifested in the acoustic pattern as a function of time, showing changes in acoustic dynamics e.g. formant slopes. The conclusion derived from the study is that articulatory dynamic aspects (e.g. speed of

tongue or lip movements), as well as articulatory static aspects (e.g. size and shape of the vocal tract), are the sources of variations in the acoustic dynamics of real speech signals.

6.3 On the Separation of Speech Signal Variances from Two Sources

Chapter 5 has described the methods which were developed to separately model the characteristics of the acoustic realisation of phonetic variation and the variance due to speaker variation. With uniform segmentation of each observation sequence into states in cepstral domain when forming the initial model, the influence of the acoustic dynamics on the variance of individual state (representing speaker variance) is able to be reasonably uniform across states in a model. With the relative distance between any two adjacent states of a hidden Markov model introduced and the same criterion of the averaged relative distance across states in a model applied, the influence of the acoustic dynamics on the speaker variance is able to be similar over different speech events modeled. If the speaker variances are the same between different speech events modeled, the resolution of the acoustic dynamics will be modeled similar between the speech events. Therefore, it becomes possible to provide an analytical comparison of a repeated performance of a given phonemic sequence between different speakers or between different physiological or emotional states of a speaker, or an analytical comparison of a repeated performance between different phonemic sequences by one speaker, in terms of the phonetic dynamism applied and the other intra-speaker variants involved.

It would be recommended to check whether the selected acoustic parameters change monotonically with time when the methods are applied to analysis of the speech signals,. In the study, it has been presumed that the relative distance between any adjacent states decreases as the number of states in the

model increases. The assumption is valid if the acoustic parameters change monotonically with time, such as the cepstral coefficients of vocalic parts of sounds "we", "high" and "how". However, the assumption seems not always to be valid for the speech signals, the acoustic parameters of which are non-monotonic with time. As an extreme case, the English word "*sauce*" has symmetric phonemic sequence. Its cepstral coefficients of the vocalic part could be symmetric or semi-symmetric. In this case, the relative distance between the states with 2-states model could be very small and less than that between the adjacent states with 3-states model. In an analysis, the starting trial of the number of states is not necessarily to be 2 and can be determined based on the prior knowledge about the speech signal before modeling.

The methods developed have paved the way for further study of the variances from different sources, which seem to be ignored in speech engineering research. The experiments have demonstrated different characteristics across index of the cepstral coefficients for the selected sound/speaker pairs between the intra-speaker repetition variance, which was modeled with $D_0^2 = 4$, and the global variance which is the combination of two types of variance: the phonetic variance and the intra-speaker repetition variance. This difference is similar to the phenomenon noted by Juang *et al.* (1987) in their experiments. Juang *et al.* presented the difference between the variance of "simulated data" (corresponding to the intra-speaker repetition variance in the thesis study) and the variance of "mixed data" (corresponding to the global variance in the thesis study) by the ratio of the former to the latter for each cepstral dimension. The ratio increases (or reciprocal of the ratio decreases as reported in this thesis study) as cepstral index increases. They interpreted this phenomenon as indicating "*the diminishing discriminating power of the higher quefrency terms*" and showing "*that the variability of higher quefrency terms are inherent artifacts of the analysis procedure, and hence less desirable in spectral similarity comparisons than lower quefrency terms*". However the results of this study suggest that the observed difference is caused by the phonetic variance in-

cluded in the variance of "mixed data" (by the phonetic variance included in the global variance in the thesis study). One should be very careful when using the method proposed by Juang *et al.* (1987) as a consequence of the analysis to compress the value of higher cepstral components and to form new measurement vectors for the speech recognition.

The differences between the intra-speaker repetition variance and the global variance in the study, have been shown in both the overall and individual quantities of cepstral coefficients. In normal circumstances, the speech signals from a speaker for training a model-based speaker recognition system cover wide a range of sounds. The variance over all training data in cepstral space can be dominated by the phonetic variation. The variance of individual codeword, in a Gaussian mixture model or a degenerated one, can be dominated by intra-speaker repetition variance. This study does not support for the use of covariance from all pooled speech materials of a speaker in place of the covariance of individual codeword as suggested by Soong *et al.* (1988).

6.4 Automatic Speaker Recognition Research from Two Perspectives

In undertaking studies on the speaker recognition and its related topics, we have clearly seen that many efforts have been made in the engineering research to introduce advanced mathematical methods and novel techniques to effectively describe a given speech data set from a speaker. This direction should be continued as it can be seen for example, that we need to find more appropriate model form in order to build up an optimal recognition system.

We also have seen the fundamental role that the scientific research plays in the speaker recognition. One example is the situation where some characteristic of the test data is not observed in the training data due to speaker variation. Rosenberg *et al.* have raised this problem in their review paper

(1991). When it happens, the assumption for our currently widely-used Bayes rule based system is broken down. The increase of the training data over different periods and different physiological or emotional states could be used to keep updating the system. However, it is believed that the scientific studies on the ways a speaker produces his speech signals and the relations between the speaker's characteristics and the speech signals can eventually help to provide a comprehensive model of a speaker, which not only accumulates all his current characteristics , but also can predict any potential changes of them.

Our current understanding of speech production and its relations to the speech signals is still at primitive stage. The understanding of sources of speaker variance and their influences on both static and dynamic speech signals is mostly at qualitative level. We have just modified our knowledge about sources of speaker variance in the acoustic dynamics of speech. It can be seen that there is a long way to go from qualitative understanding to quantitative description in the scientific research. In the thesis study, an attempt has been made to develop a method for the scientific research to be able to conduct quantitative analysis of the speech signals in terms of acoustic dynamics realised and speaker variance involved . The efforts made by several generations of the speech researchers, and the methods and techniques developed in the engineering research, will help in progress of the scientific research of automatic speaker recognition.

6.5 Contributions of the Thesis

The contributions of the thesis work to automatic speaker recognition and, in general, to speech science and engineering, are summarised as follows,

- The mathematical relations between VQ model and GM model, which are widely used in the speaker recognition systems, have been clearly revealed, showing that GM model degenerates to VQ model when the three constraints are imposed on it: 1) non-overlapping between components,

2) equal mixture weights, 3) identity of covariance matrix of each component . The experimental results have demonstrated that those three constraints on GM model are not suitable to accurately model speakers for the speaker recognition.

- A method has been developed to re-adjust the parameters of speakers' models in order to improve correctness of the ranking order of probabilities between the models, hence improve the performance of a model-based speaker recognition system. Correct ranking order of the probabilities between models is essential for the optimal performance of Bayesian rule based recognition tasks. It has been demonstrated that the improvement can be realised when the models estimated by the maximum likelihood method do not correctly represent the real distributions.

- The analysis of relations of the models further re-estimated by the maximum mutual information method (the MMIE) to the models estimated by the maximum likelihood method (the MLE) has shown that the re-estimated models will remain unchanged if they are initially estimated by the MLE and at least one model has single maximum value. It has been demonstrated that the models with greater value of the statistics defined in the MLE and the MMIE do not necessarily lead to better identification performance if the model form is not completely correct.

- The thesis study has revealed another perspective of the nonlinear relationship between the articulatory parameters and the acoustic parameters of the vocal tract. It has been demonstrated that the size and shape of the vocal tract affect the values of formants non-uniformly along the particular regions the articulators pass through when producing sounds. When the articulators move through those regions with time, this effect will be manifested in the acoustic pattern as a function of time, showing changes in acoustic dynamics e.g. formant slopes. The conclusion is that the size and shape of the vocal tract, as well as speed of tongue or

lip movements, are the sources of variances in acoustic dynamics of the speech.

- A method has been developed to model a dynamical speech event such that the resolution of the acoustic dynamics and its influence on the acoustic speaker variance accompanied can be quantitatively controlled. This method can be used to analytically compare a repeated performance of a given phonemic sequence between different speakers or between different physiological or emotional states of a speaker, or a repeated performance between different phonemic sequences by one speaker.

- With the method mentioned above, the acoustic speaker variance manifested in the acoustic dynamics of the speech can be extracted. The qualitative and quantitative differences in LPC_cepstral space between the intra-speaker repetition variance and the mixed variance, which includes both intra-speaker and phonetic variance, have been revealed in the study. A mixed variance measure is often seen to be used in the engineering research. The study suggests that the different kinds of sources should be clarified before applying the acoustic measures in practical systems.

Appendix A

Model Parameters for the Simulated Sounds

This appendix contains the tables, each of which lists values of the static parameters and boundary conditions of the time-varying parameters of the vocal tract model for one simulation of each of three dynamic segments /aI/, /aU/ and /wi/. They have been used in the study on the sources of inter- and intra- speaker variability in the acoustic dynamics of speech (see Chapter 4).

On each page, the table labeled as (b) lists values of the parameters for benchmark simulation of one sound. The tables labeled as (a) and (c) list values of the parameters for a simulation of one sound with one parameter perturbed. The perturbed parameter and its selected value are shown in bold type style of font in the table.

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$A_l(R_l)$	8 (1.6)
L	**16.5**	$A_m(R_m)$	8 (1.6)
$\mathbf{X_c(0)}$	**4.4**	$A_c(0)(R_c(0))$	.65(.455)
$\mathbf{X_c(1)}$	**9.9**	$A_c(.5)(R_c(.5))$	3.14 (1)
		$A_c(1)(R_c(1))$	.65(.455)

(a)

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$A_l(R_l)$	8 (1.6)
L	17.6	$A_m(R_m)$	8 (1.6)
$X_c(0)$	5.5	$A_c(0)(R_c(0))$	.65 (.455)
$X_c(1)$	11	$A_c(0.5)(R_c(0.5))$	3.14 (1.)
		$A_c(1)(R_c(1))$	.65 (.455)

(b)

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$A_l(R_l)$	8 (1.6)
L	**18.7**	$A_m(R_m)$	8 (1.6)
$\mathbf{X_c(0)}$	**6.6**	$A_c(0)(R_c(0))$	.65 (.455)
$\mathbf{X_c(1)}$	**12.1**	$A_c(.5)(R_c(.5))$	3.14 (1)
		$A_c(1)(R_c(1))$	.65 (.455)

(c)

Table A.1: Parameters of the model for /aI/, (a): $L = 16.5$, (b): $L = 17.6$, (c): $L = 18.7$.

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$A_l(R_l)$	8 (1.6)
L	17.6	$A_m(R_m)$	8 (1.6)
$X_c(0)$	5.5	$\mathbf{A_c(0)(R_c(0))}$	**.289 (.303)**
$X_c(1)$	11	$\mathbf{A_c(.5)(R_c(.5))}$	**2.26 (.848)**
		$\mathbf{A_c(1)(R_c(1))}$	**.289 (.303)**

(a)

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$A_l(R_l)$	8 (1.6)
L	17.6	$A_m(R_m)$	8 (1.6)
$X_c(0)$	5.5	$A_c(0)(R_c(0))$	.65 (.455)
$X_c(1)$	11	$A_c(0.5)(R_c(0.5))$	3.14 (1.)
		$A_c(1)(R_c(1))$	.65 (.455)

(b)

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$A_l(R_l)$	8 (1.6)
L	17.6	$A_m(R_m)$	8 (1.6)
$X_c(0)$	5.5	$\mathbf{A_c(0)(R_c(0))}$	**1.16 (.607)**
$X_c(1)$	11	$\mathbf{A_c(.5)(R_c(.5))}$	**4.17 (1.152)**
		$\mathbf{A_c(1)(R_c(1))}$	**1.16 (.607)**

(c)

Table A.2: Parameters of the model for /aI/, (a): $A_c(0) = .289$, (b): $A_c(0) = .65$, (c): $A_c(0) = 1.16$.

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$\mathbf{A_l(R_l)}$	**14.2 (2.13)**
L	17.6	$\mathbf{A_m(R_m)}$	**14.2 (2.13)**
$X_c(0)$	5.5	$A_c(0)(R_c(0))$	.65 (.455)
$X_c(1)$	11	$A_c(0.5)(R_c(0.5))$	3.14 (1.)
		$A_c(1)(R_c(1))$	.65 (.455)

(a)

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$A_l(R_l)$	8 (1.6)
L	17.6	$A_m(R_m)$	8 (1.6)
$X_c(0)$	5.5	$A_c(0)(R_c(0))$	.65 (.455)
$X_c(1)$	11	$A_c(0.5)(R_c(0.5))$	3.14 (1.)
		$A_c(1)(R_c(1))$	.65 (.455)

(b)

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$\mathbf{A_l(R_l)}$	**3.57 (1.07)**
L	17.6	$\mathbf{A_m(R_m)}$	**3.57 (1.07)**
$X_c(0)$	5.5	$A_c(0)(R_c(0))$	.65 (.455)
$X_c(1)$	11	$A_c(0.5)(R_c(0.5))$	3.14 (1.)
		$A_c(1)(R_c(1))$	.65 (.455)

(c)

Table A.3: Parameters of the model for /aI/, (a): $A_m = 14.2$, (b): $A_m = 8.$, (c): $A_m = 3.57$.

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$A_m(R_m)$	8 (1.6)
L	**16.5**	$A_c(R_c)$	.65 (.455)
$\mathbf{X_c(0)}$	**4.4**	$A_l(0)(R_l(0))$	8 (1.6)
$\mathbf{X_c(1)}$	**7.4**	$A_l(1)(R_l(1))$	.16 (.226)

(a)

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$A_m(R_m)$	8 (1.6)
L	17.6	$A_c(R_c)$	.65 (.455)
$X_c(0)$	5.5	$A_l(0)(R_l(0))$	8 (1.6)
$X_c(1)$	8.5	$A_l(1)(R_l(1))$	.16 (.226)

(b)

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$A_m(R_m)$	8 (1.6)
L	**18.7**	$A_c(R_c)$	.65 (.455)
$\mathbf{X_c(0)}$	**6.6**	$A_l(0)(R_l(0))$	8 (1.6)
$\mathbf{X_c(1)}$	**9.6**	$A_l(1)(R_l(1))$	.16 (.226)

(c)

Table A.4: Parameters of the model for /aU/, (a): $L = 16.5$, (b): $L = 17.6$, (c): $L = 18.7$.

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$\mathbf{A_c(R_c)}$	**.289 (.303)**
L	17.6	$A_m(R_m)$	8 (1.6)
$X_c(0)$	5.5	$A_l(0)(R_l(0))$	8 (1.6)
$X_c(1)$	8.5	$A_l(1)(R_l(1)$	.16 (.226)

(a)

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$A_m(R_m)$	8 (1.6)
L	17.6	$A_c(R_c)$	.65 (.455)
$X_c(0)$	5.5	$A_l(0)(R_l(0))$	8 (1.6)
$X_c(1)$	8.5	$A_l(1)(R_l(1))$	.16 (.226)

(b)

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$\mathbf{A_c(R_c)}$	**1.16 (.607)**
L	17.6	$A_m(R_m)$	8 (1.6)
$X_c(0)$	5.5	$A_l(0)(R_l(0))$	8 (1.6)
$X_c(1)$	8.5	$A_l(1)(R_l(1)$	.16 (.226)

(c)

Table A.5: Parameters of the model for /aU/, (a): $A_c = .289$, (b): $A_c = 0.65$, (c): $A_c = 1.16$.

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$A_c(R_c)$	.65 (.455)
L	17.6	$\mathbf{A_m(R_m)}$	**14.2 (2.13)**
$X_c(0)$	5.5	$A_l(0)(R_l(0))$	8 (1.6)
$X_c(1)$	8.5	$A_l(1)(R_l(1)$	.16 (.226)

(a)

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$A_m(R_m)$	8 (1.6)
L	17.6	$A_c(R_c)$	.65 (.455)
$X_c(0)$	5.5	$A_l(0)(R_l(0))$	8 (1.6)
$X_c(1)$	8.5	$A_l(1)(R_l(1))$	.16 (.226)

(b)

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$A_c(R_c)$	.65 (.455)
L	17.6	$\mathbf{A_m(R_m)}$	**3.57 (1.07)**
$X_c(0)$	5.5	$A_l(0)(R_l(0))$	8 (1.6)
$X_c(1)$	8.5	$A_l(1)(R_l(1)$	.16 (.226)

(c)

Table A.6: Parameters of the model for /aU/, (a): $A_m = 14.2$, (b): $A_m = 8.$, (c): $A_m = 3.57$.

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$A_c(R_c)$	.65 (.455)
L	**16.5**	$A_m(R_m)$	8 (1.6)
$\mathbf{X_c(0)}$	**7.9**	$A_l(0)(R_l(0))$	.16 (.226)
$\mathbf{X_c(1)}$	**11.4**	$A_l(1)(R_l(1))$	8. (1.6)

(a)

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$A_c(R_c)$	.65 (.455)
L	17.6	$A_m(R_m)$	8 (1.6)
$X_c(0)$	9	$A_l(0)(R_l(0))$	.16(.226)
$X_c(1)$	12.5	$A_l(1)(R_l(1)$	8.(1.6)

(b)

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$A_c(R_c)$	.65 (.455)
L	**18.7**	$A_m(R_m)$	8 (1.6)
$\mathbf{X_c(0)}$	**10.1**	$A_l(0)(R_l(0))$	.16(.226)
$\mathbf{X_c(1)}$	**13.6**	$A_l(1)(R_l(1))$	8.(1.6)

(c)

Table A.7: Parameters of the model for /wi/, (a): $L = 16.5$, (b): $L = 17.6$, (c): $L = 18.7$.

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$A_c(R_c)$	.65 (.455)
L	17.6	$\mathbf{A_m(R_m)}$	**14.2 (2.13)**
$X_c(0)$	9.	$A_l(0)(R_l(0))$	.16 (.226)
$X_c(1)$	12.5	$A_l(1)(R_l(1))$	8.(1.6)

(a)

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$A_c(R_c)$	.65 (.455)
L	17.6	$A_m(R_m)$	8 (1.6)
$X_c(0)$	9	$A_l(0)(R_l(0))$	.16(.226)
$X_c(1)$	12.5	$A_l(1)(R_l(1)$	8.(1.6)

(b)

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$A_c(R_c)$	.65 (.455)
L	17.6	$\mathbf{A_m(R_m)}$	**3.57 (1.07)**
$X_c(0)$	9.	$A_l(0)(R_l(0))$	.16 (.226)
$X_c(1)$	12.5	$A_l(1)(R_l(1))$	8.(1.6)

(c)

Table A.8: Parameters of the model for /wi/, (a): $A_m = 14.2$, (b): $A_m = 8.$, (c): $A_m = 3.57$.

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$\mathbf{A_c(R_c)}$	**.289 (.303)**
L	17.6	$A_m(R_m)$	8 (1.6)
$X_c(0)$	9.	$A_l(0)(R_l(0))$	8 (1.6)
$X_c(1)$	12.5	$A_l(1)(R_l(1))$	.16 (.226)

(a)

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$A_c(R_c)$	.65 (.455)
L	17.6	$A_m(R_m)$	8 (1.6)
$X_c(0)$	9	$A_l(0)(R_l(0))$	.16(.226)
$X_c(1)$	12.5	$A_l(1)(R_l(1)$	8.(1.6)

(b)

Parameter	Value	Parameter	Value
L_g	2.2	A_g	1.5
L_m	1.1	$\mathbf{A_c(R_c)}$	**1.16 (.607)**
L	17.6	$A_m(R_m)$	8 (1.6)
$X_c(0)$	9.	$A_l(0)(R_l(0))$	8 (1.6)
$X_c(1)$	12.5	$A_l(1)(R_l(1))$	.16 (.226)

(c)

Table A.9: Parameters of the model for /wi/, (a): $A_c = .289$, (b): $A_c = .65$, (c): $A_c = 1.16$.

Appendix B

The Figures of the Global Variance and Intra-speaker Repetition Variance

This appendix contains figures of the global variance and intra-speaker repetition variance over 12 LPC cepstral coefficients for the vocalic part of word "we", "how" and "high" of the Monosyllabic Word database by all the speakers except speaker AH (the figures for his data have been shown in Section 5.3).

The intra-speaker repetition variance was derived based on the variance averaged over all states in the model which has been developed with $D_0^2 = 4$ (see Chaper 5). The global variance was derived based on all observations of the sound/speaker pair. This variance is a combination of phonetic variance and intra-speaker repetition variance.

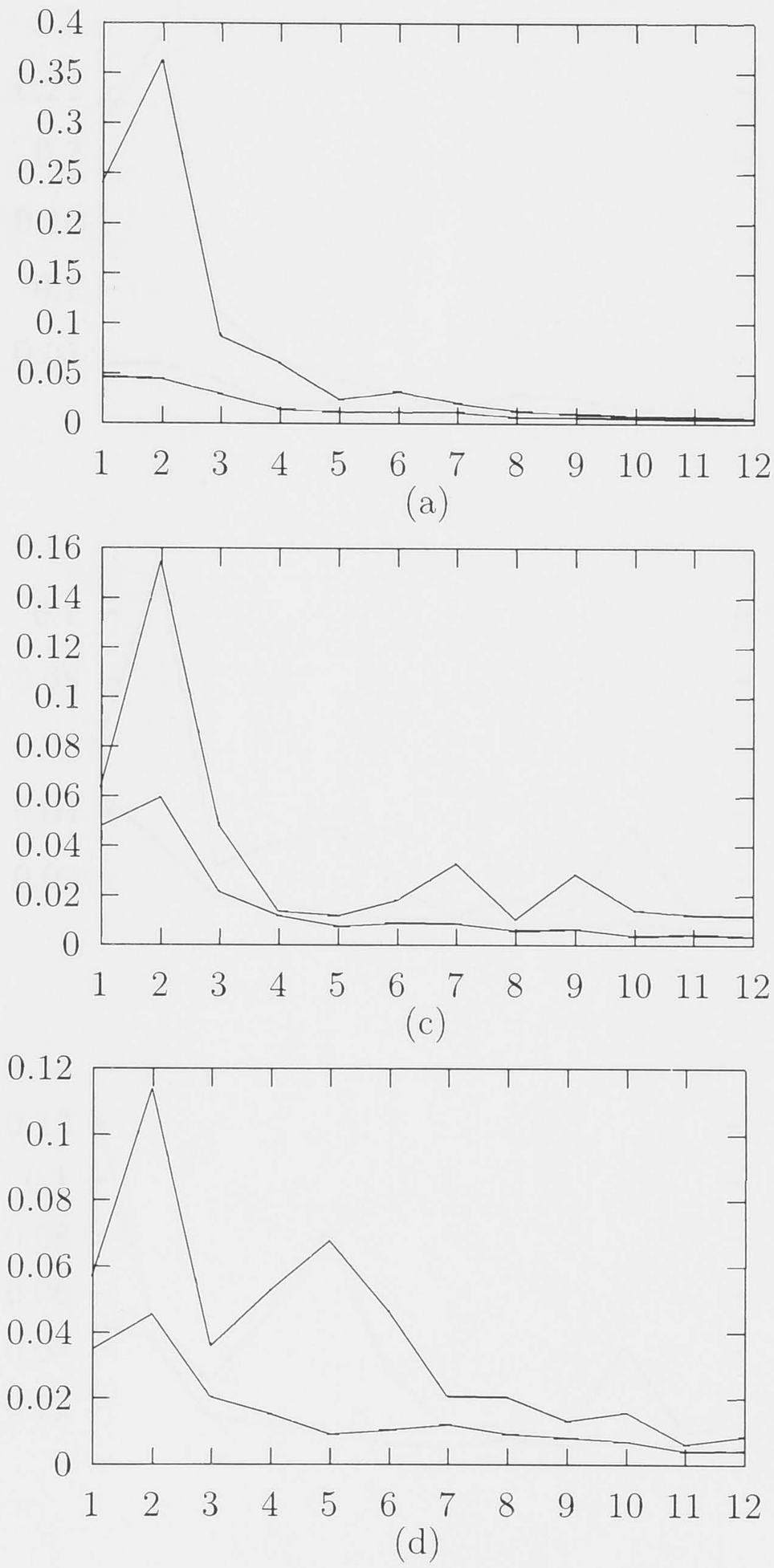


Figure B-1: The global variance and intra-speaker repetition variance against each of 12 LPC cepstral coefficients; (a): "we", (c): "how", (d): "high" by speaker BM. The upper curve is the global variance. The index of the cepstral coefficients goes from 1 to 12.

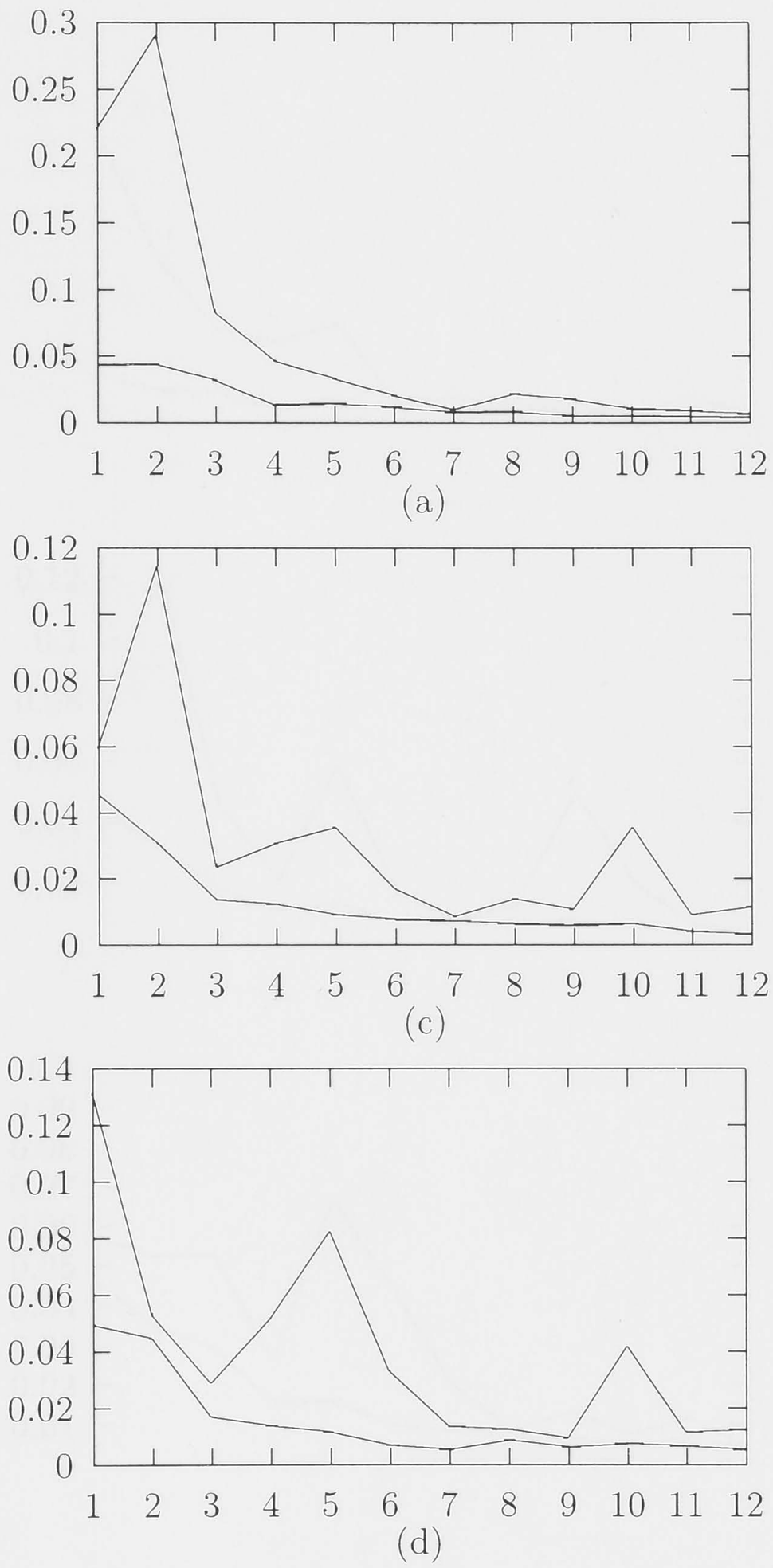


Figure B-2: The global variance and intra-speaker repetition variance against each of 12 LPC cepstral coefficients; (a): "we", (c): "how", (d): "high" by speaker DD. The upper curve is the global variance. The index of the cepstral coefficients goes from 1 to 12.

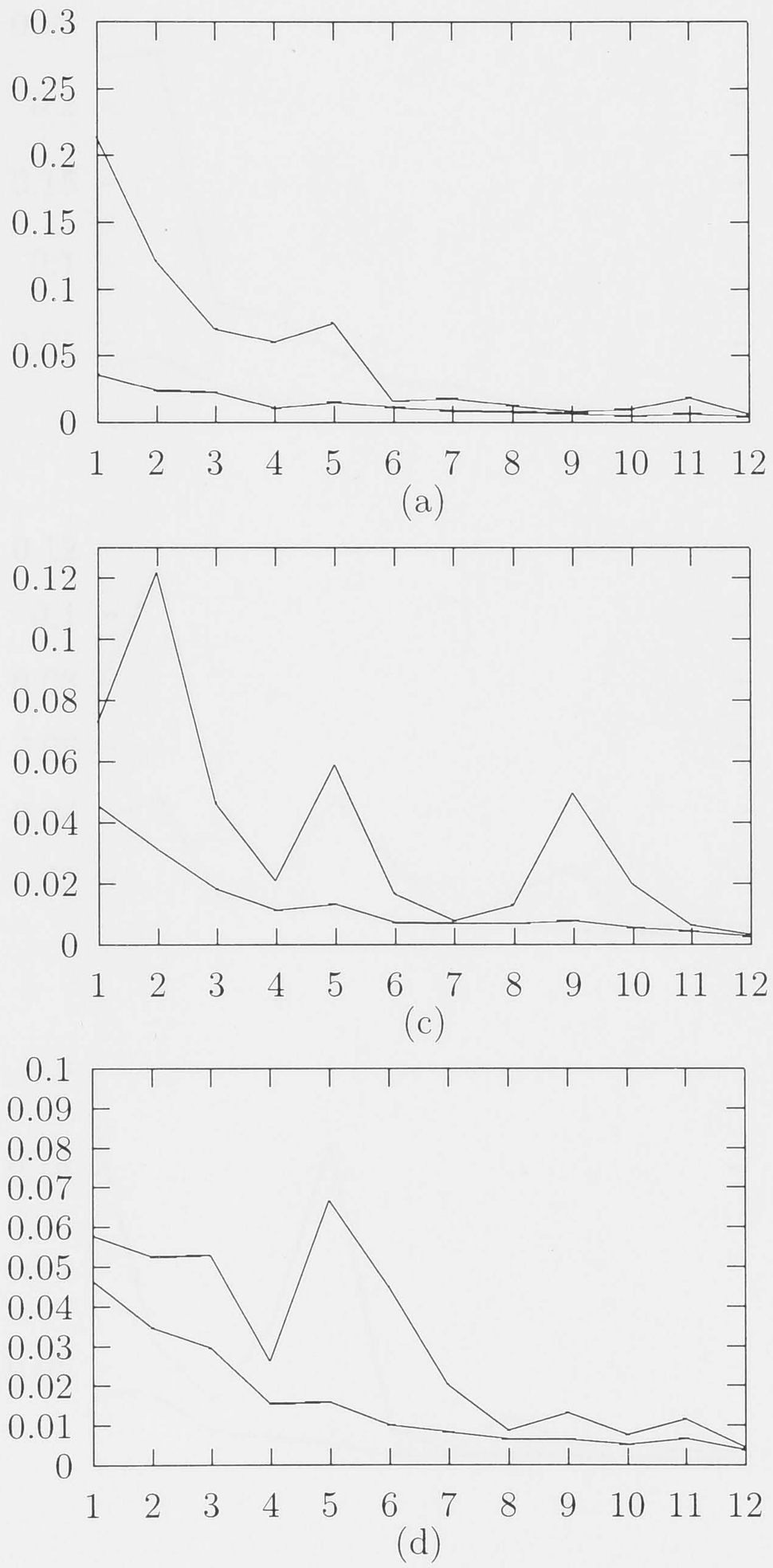


Figure B-3: The global variance and intra-speaker repetition variance against each of 12 LPC cepstral coefficients; (a): "we", (c): "how", (d): "high" by speaker GC. The upper curve is the global variance. The index of the cepstral coefficients goes from 1 to 12.

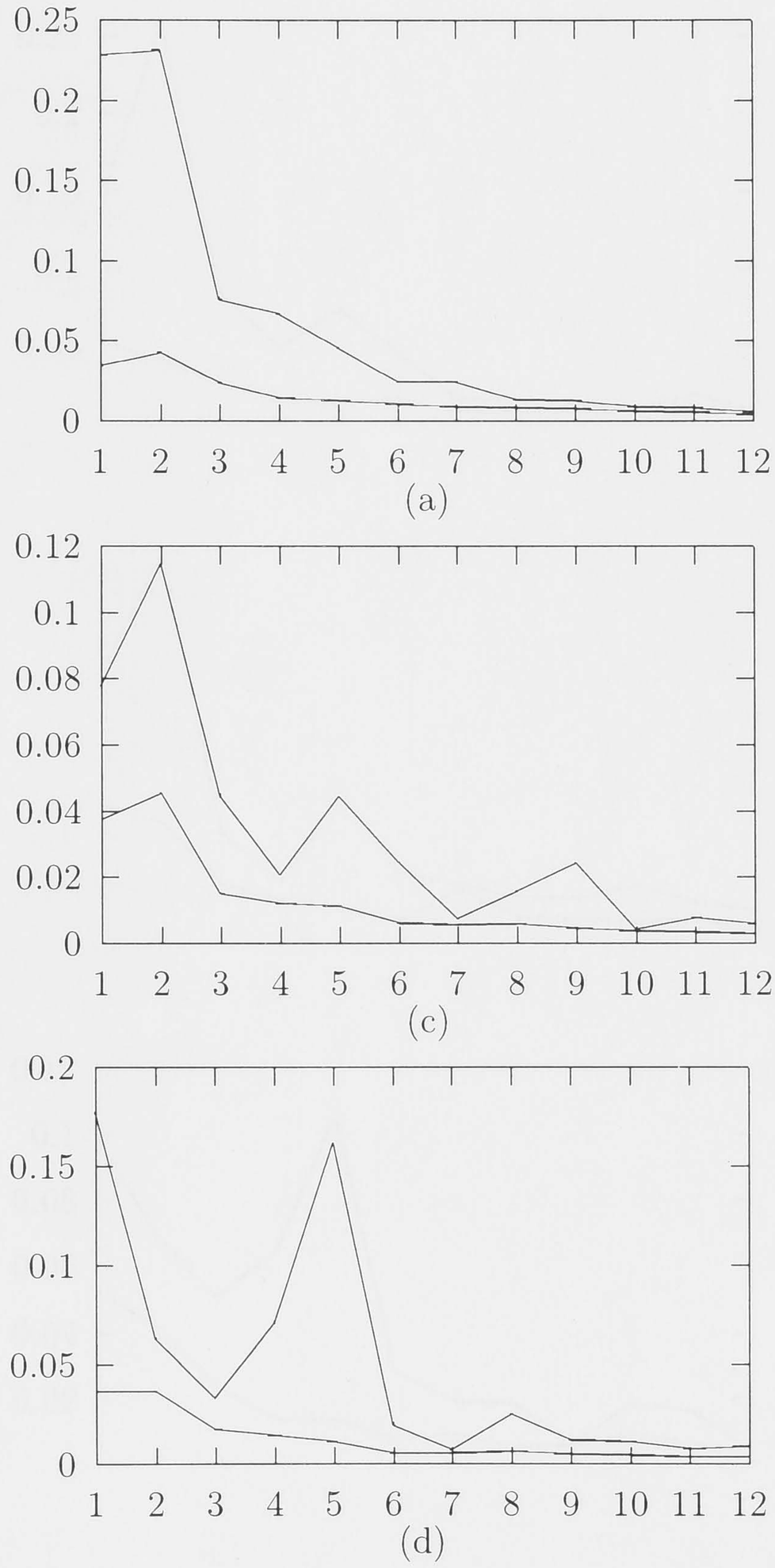


Figure B-4: The global variance and intra-speaker repetition variance against each of 12 LPC cepstral coefficients; (a): "we", (c): "how", (d): "high" by speaker IM. The upper curve is the global variance. The index of the cepstral coefficients goes from 1 to 12.

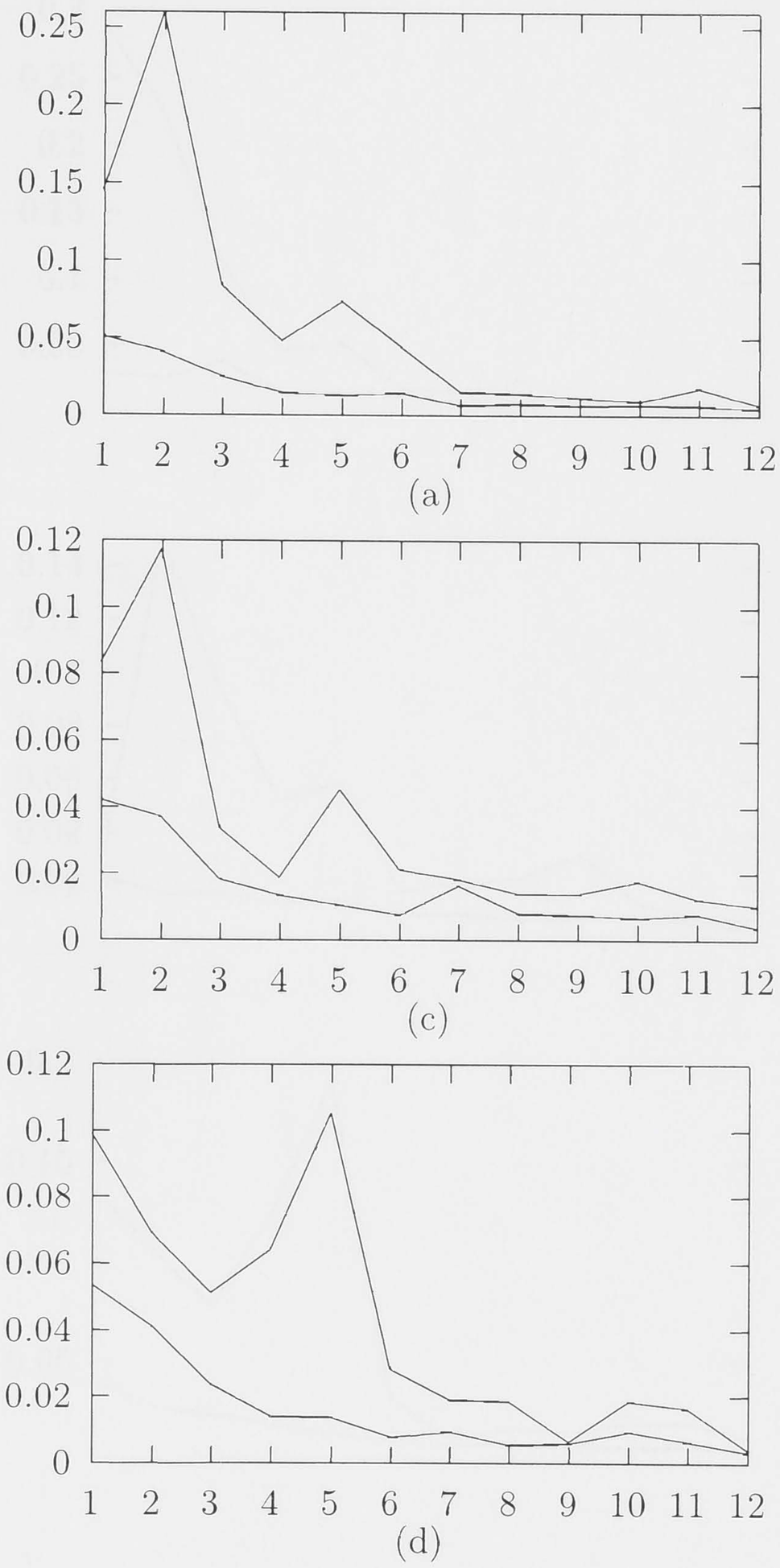


Figure B-5: The global variance and intra-speaker repetition variance against each of 12 LPC cepstral coefficients; (a): "we", (c): "how", (d): "high" by speaker JW. The upper curve is the global variance. The index of the cepstral coefficients goes from 1 to 12.

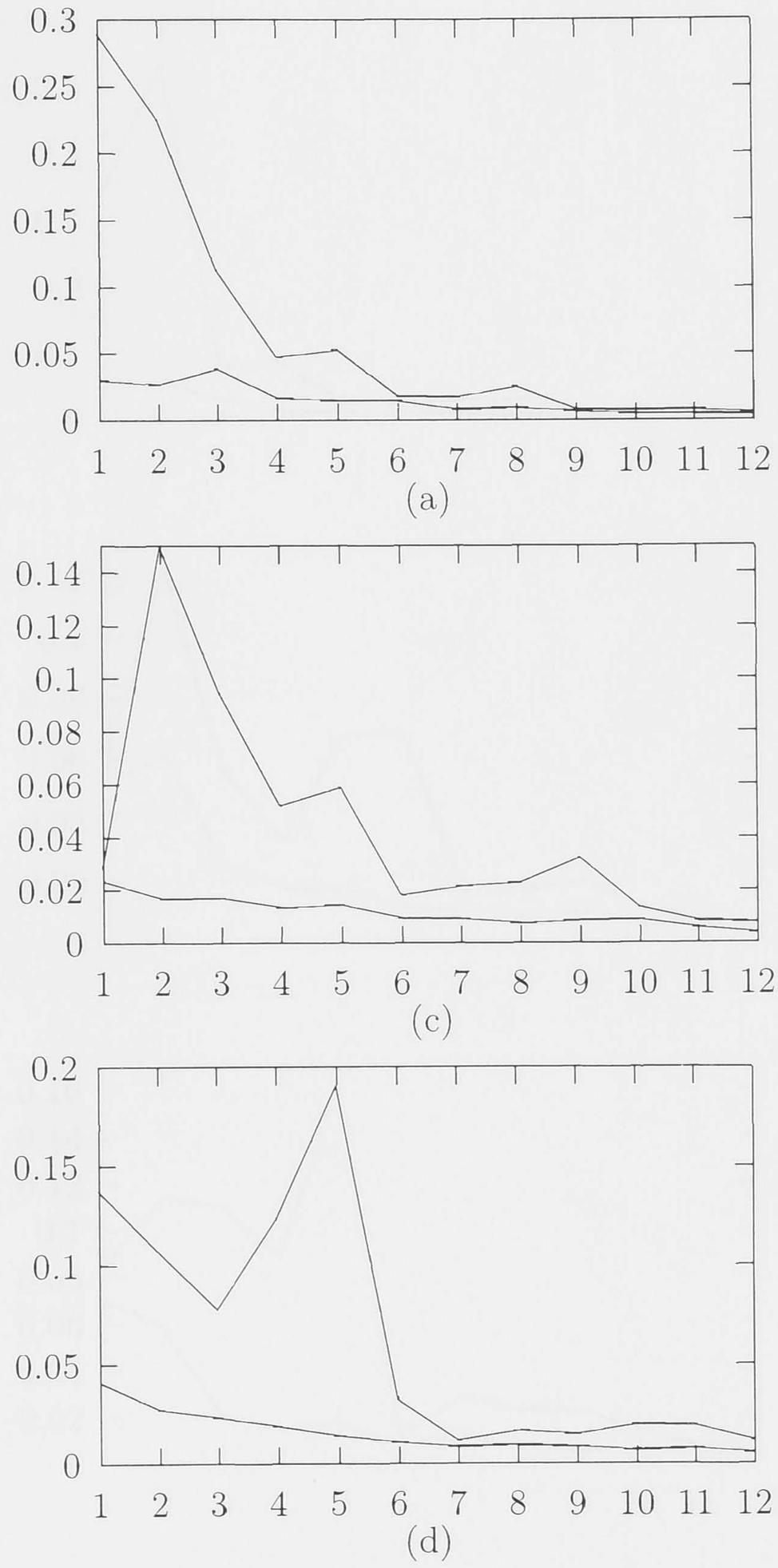


Figure B-6: The global variance and intra-speaker repetition variance against each of 12 LPC cepstral coefficients; (a): "we", (c): "how", (d): "high" by speaker KR. The upper curve is the global variance. The index of the cepstral coefficients goes from 1 to 12.

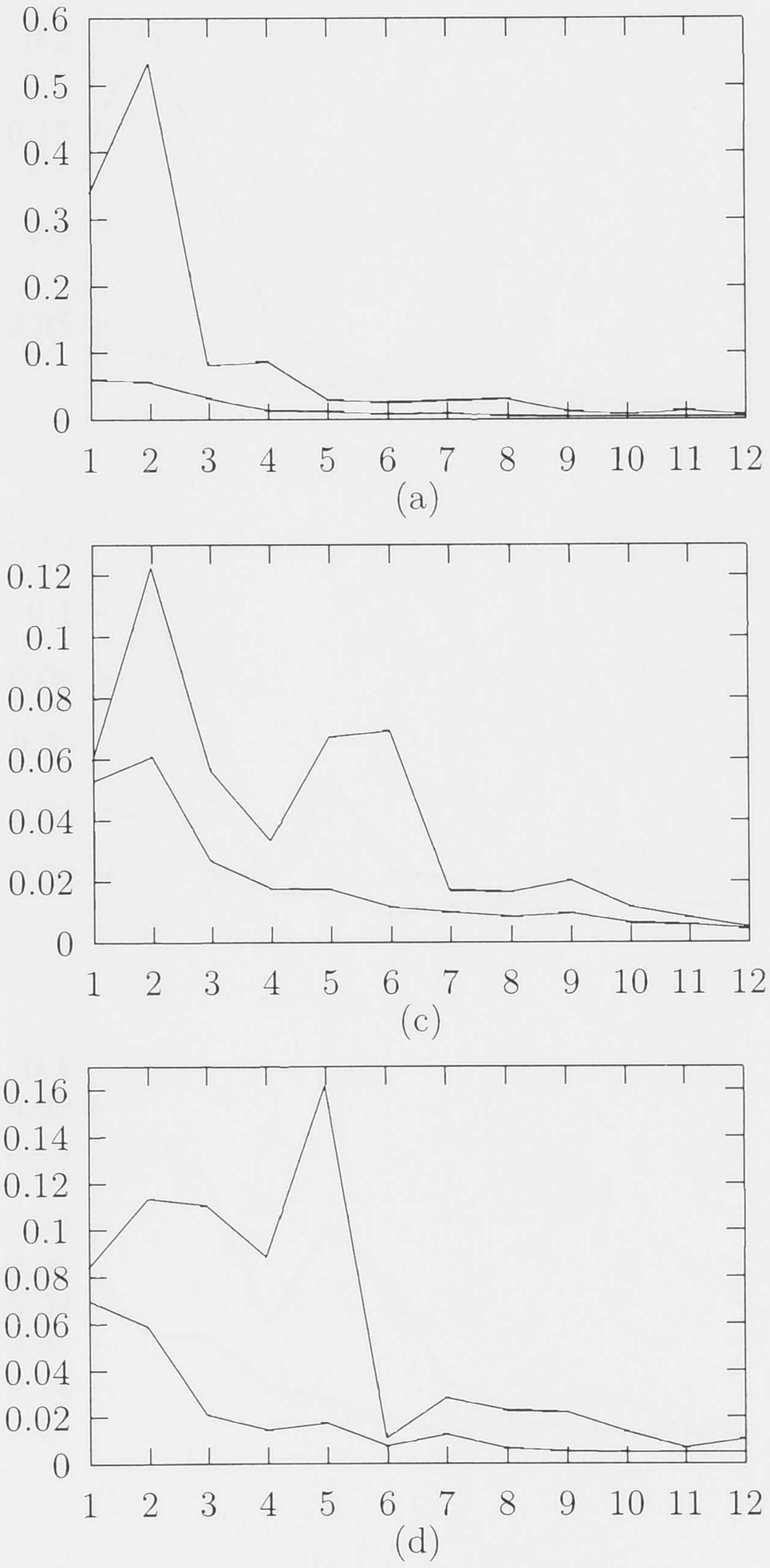


Figure B-7: The global variance and intra-speaker repetition variance against each of 12 LPC cepstral coefficients; (a): "we", (c): "how", (d): "high" by speaker NF. The upper curve is the global variance. The index of the cepstral coefficients goes from 1 to 12.

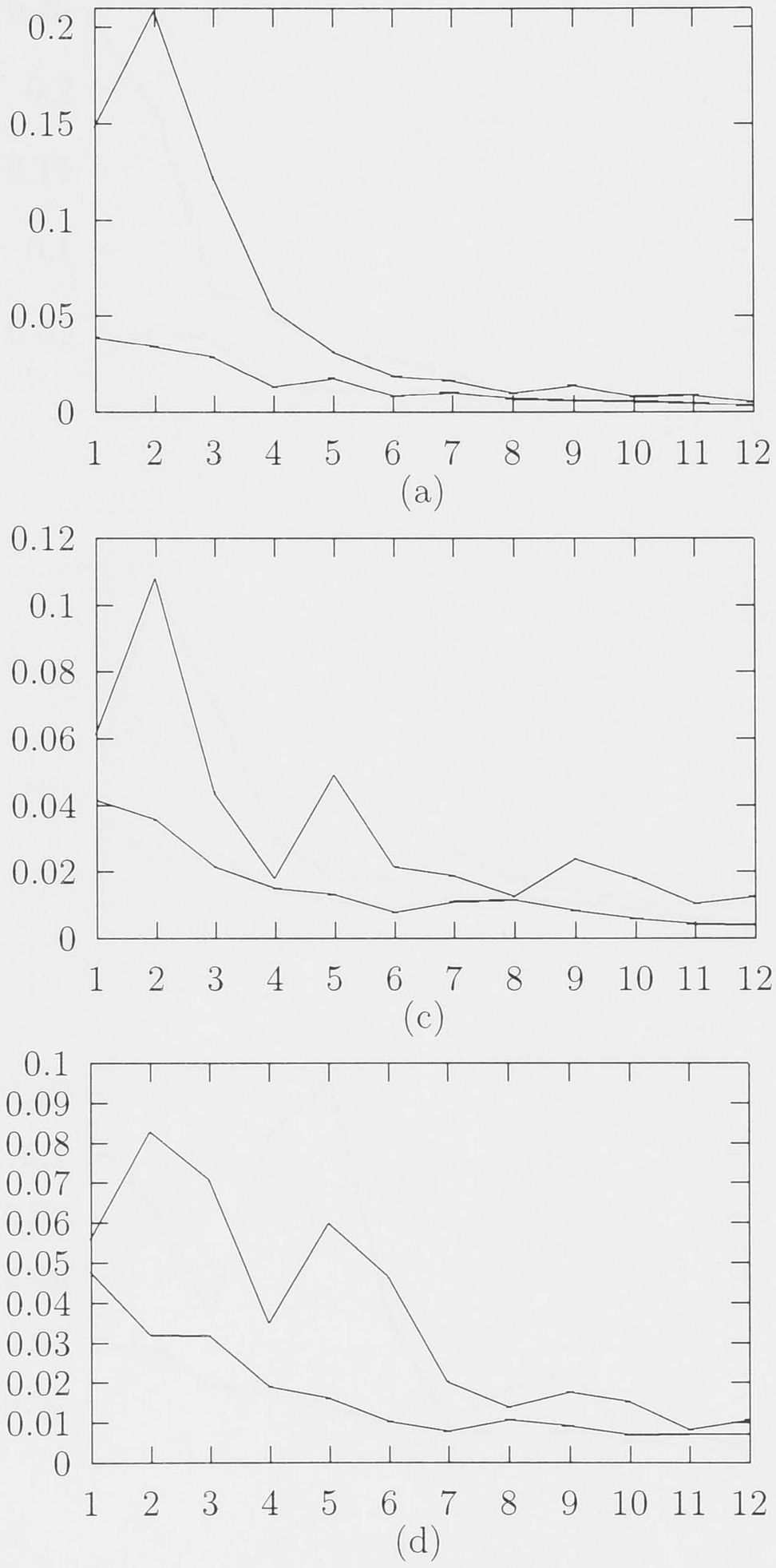


Figure B-8: The global variance and intra-speaker repetition variance against each of 12 LPC cepstral coefficients; (a): "we", (c): "how", (d): "high" by speaker SB. The upper curve is the global variance. The index of the cepstral coefficients goes from 1 to 12.

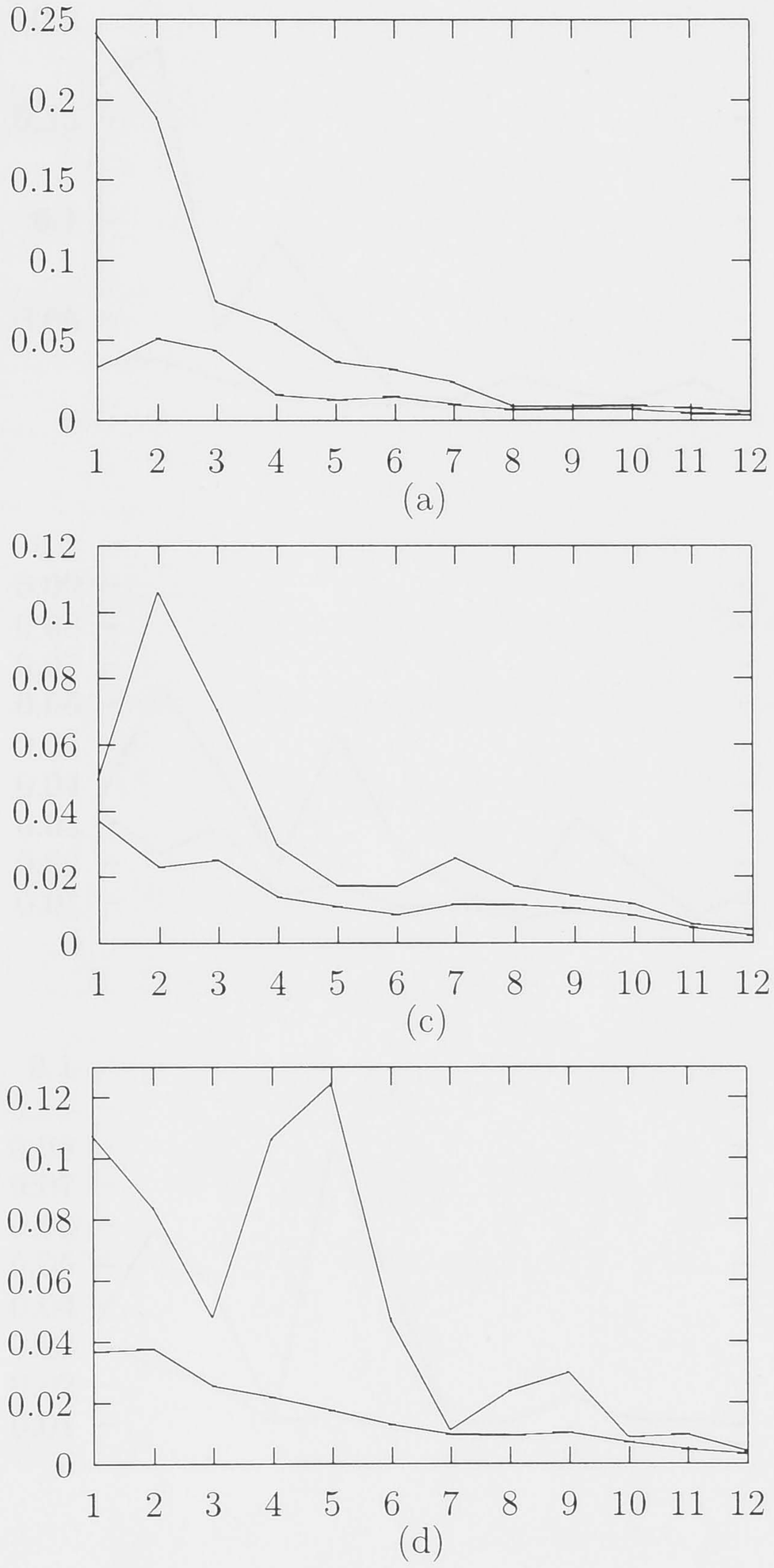


Figure B-9: The global variance and intra-speaker repetition variance against each of 12 LPC cepstral coefficients; (a): "we", (c): "how", (d): "high" by speaker TB. The upper curve is the global variance. The index of the cepstral coefficients goes from 1 to 12.

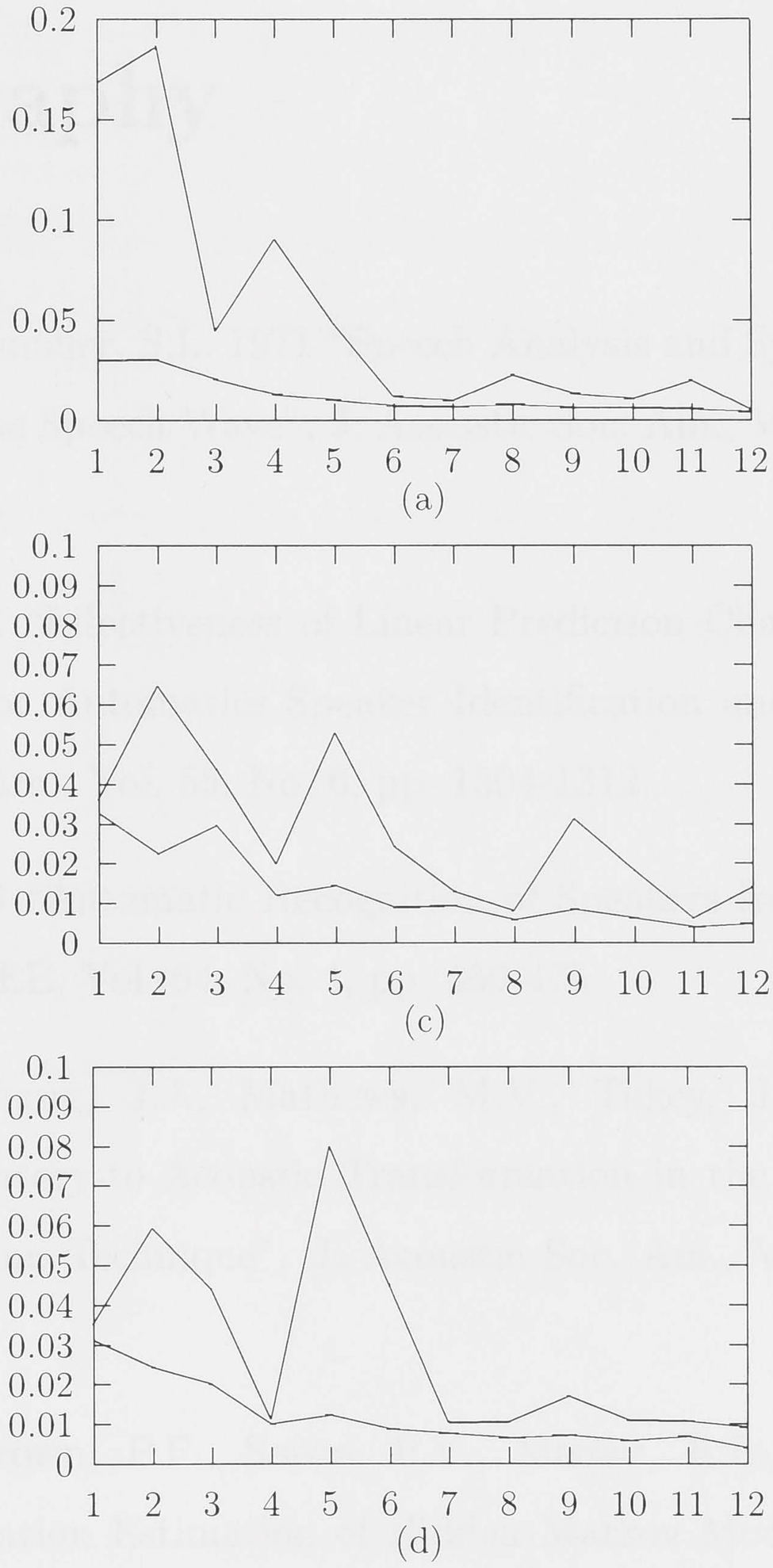


Figure B-10: The global variance and intra-speaker repetition variance against each of 12 LPC cepstral coefficients; (a): "we", (c): "how", (d): "high" by speaker WB. The upper curve is the global variance. The index of the cepstral coefficients goes from 1 to 12.

Bibliography

[1] Atal, B.S. & Hanauer, S.L. 1971 "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", J. Acoustic Soc. Am., Vol. 50, No. 2 (part 2), pp. 637-655.

[2] Atal, B.S. 1974 "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatics Speaker Identification and Verification", J. Acoustic Soc. Am., Vol. 55, No. 6, pp. 1304-1312.

[3] Atal, B.S. 1976 "Automatic Recognition of Speakers from Their Voices", Proc. of the IEEE, Vol. 64, No. 4, pp. 460-475.

[4] Atal, B.S., Chang, J.J., Mathews, M.V., Tukey, J.W. 1978 "Inversion of Articulatory-to-Acoustic Transformation in the Vocal Tract by a Computer-sorting Technique", J. Acoustic Soc. Am., Vol. 63, No. 5, pp. 1535-1555.

[5] Bahl, L.R., Brown, P.F., Souza, P.V., Mercer, R.L. 1986 "Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition", Proc. of ICASSP'86, pp. 49-52.

[6] Bahl, L.R, Brown, P.F., Souza, P.V., Mercer, R.L. 1988 "A New Algorithm for the Estimation of Hidden Markov Model Parameters", Proc. of ICASSP'88, pp. 493-496.

[7] Bahl, L.R., Brown, P.F., Souza, P.V., Mercer, R.L. 1990 "Speech Recognition with Continuous-Parameter Hidden Markov Model" in Readings in

Speech Recognition edited by Waibel, A.& Lee, K., Morgan Kaufmann Publisher, Inc.: San Mateo, California, pp. 332-339.

[8] Clermont, F. 1991 "Formant-contour Models of Diphthongs: a Study in Acoustic Phonetics and Computer Modeling of Speech", Ph.D. thesis, The Australia National University.

[9] Clermont, F. 1993 "Spectro-temporal Description of Diphthongs in F1-F2-F3 Space", Speech Communication, Vol. 13, pp. 377-390.

[10] Coker, C. H. 1976 "A Model of Articulatory Dynamics and Control", Proc. of the IEEE, Vol. 64, No. 4, April, pp. 452-460.

[11] Croot, K., Fletcher, J., Harrington, J. 1992 "Levels of Segmentation and Labelling in the Australian National Database Of Spoken Language", Proc. 4th Australian International Conference on Speech Science and Technology, Brisbane, pp.86-90.

[12] Doddington, G. 1985 "Speaker Recognition - Identifying People by Their Voice", Proc. of the IEEE, Vol. 73, No. 11, pp. 1651-1664.

[13] Eatock, J. P. & Mason, J. S. 1994 "A Quantitative Assessment of the Relative Speaker Discriminating Properties of Phonemes", Proc. of ICASSP'94, pp. 133-136.

[14] Ephraim, Y., Dembo, A. & Rabiner, L.R. 1989 "A Minimum Discrimination Information Approach for Hidden Markov Modeling", IEEE Trans. on Information Theory, Vol. 35, No. 5, pp. 1001-1013.

[15] Ephraim, Y. & Rabiner, L.R. 1990 "On the Relations Between Modeling Approaches for Speech Recognition", IEEE Trans. on Information Theory, Vol. 36, No. 2, pp. 372-380.

[16] Everitt, B. S. & Hand, D. J. 1981 "Finite Mixture Distributions", Chapman and Hall, New York.

[17] Fant, G. 1960 "Acoustic Theory of Speech Production", Mouton & Co. N.V., Publishers, the Hague.

[18] Fant, G. 1975 "Vocal-tract Area and Length Perturbation", STL-QPSR 4, Speech Transmission Laboratories, K. T. H., Stockholm, pp. 1-14.

[19] Forney, G.D. 1973 "The Viterbi Algorithm", Proc. of IEEE, Vol. 61, pp. 268–278.

[20] Furui, S. 1994 "An Overview of Speaker Recognition Technology", Proc. of ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, pp. 1–9.

[21] Garvin, P.L. and Ladefoged, P. 1963 "Speaker Identification and Message Identification in Speech Recognition", Phonetica, Vol. 9, pp. 193-199.

[22] Gray, A.H. and Markel, J.D. 1976 "Distance Measure for Speech Processing", IEEE ASSP Vol. 24, No. 5, pp. 380-391.

[23] Hartigan, J. A. 1975 "Clustering Algorithms"', Wiley: New York.

[24] Hieronymus, J., Alexander, H., Bennett, C., Cohen, I., Davies, D., Dalby, J., Laver, J., Barry, W., Fourcin, A. & Well, J. 1990 "Proposed Speech Segmentation Criteria for the SCRIBE Project", SCRIBE-Project Report.

[25] Juang, B.H, Rabiner, L.R. and Wilpon, J.G. 1987 "On the Use of Bandpass Liftering in Speech Recognition", IEEE Tran. on ASSP, Vol. 35, No. 7, pp. 947-953.

[26] Juang, B.H., Rabiner, L.R. 1990 "The Segmental K-means Algorithm for Estimating Parameters of Hidden Markov Models", IEEE Trans on ASSP, Vol. 38, No. 9, pp. 1639-1641.

[27] Kent, R.D. 1972 "Some Considerations in the Cinefluorographic Analysis of Tongue Movements during Speech", Phonetica Vol. 26, pp. 16-32.

[28] Kent, R.D., Moll, K.L. 1972 "Tongue Body Articulation during Vowel and Diphthong Gestures", Folia Phoniatrica, Vol. 24, pp. 278 - 300.

[29] Lee, K.F. 1989 "Automatic Speech Recognition, the Development of the SPHINX System", Kluwer Academic Publishers.

[30] Levinson, S.E. 1986 "Continuously Variable Duration Hidden Markov Models for Automatic Speech Recognition", Computer Speech and Language, Vol. 1, pp. 29-45.

[31] Linde, Y., Buzo, A., Gray, R.M. 1980 "An Algorithm for Vector Quantization", IEEE Trans. on Commun., Vol. 28, pp.84-95.

[32] Liu, C. S., Lee, C. H., Juang, B. H., Rosenberg, A. E. 1994 "Speaker Recognition Based on Minimum Error Discriminative Training", Proc. of ICASSP'94, pp. 325-328.

[33] Ljolje, A., Ephraim, Y., Rabiner, L.R. 1990 "Estimation of Hidden Markov Model Parameters by Minimizing Empirical Error Rate", Proc. of ICASSP'90, pp. 709-712.

[34] Markel, J.D. & Gray, A.H. 1976 "Linear Prediction of Speech", Springer-Verlag: Berlin Heidelberg, New York.

[35] Matsui, T. and Furui, S. 1991 "A Text Independent Speaker Recognition Method Robust against Utterance Variation", Proc. of ICASSP, pp. 377-380.

[36] Matsui, T. & Furui, S. 1994 "Comparison of Text Independent Speaker Recognition Methods Using VQ Distortion and Discrete/Continuous HMM", IEEE Trans. on Speech & Audio Processing Vol. 2, No. 3, pp. 456-459. Also see Proc. of ICASSP'92, II-157.

[37] Mermelstein, P. 1973 "Articulatory Model for the Study of Speech Production", The Journal of the Acoustic Society of America, Vol. 53, No. 4, pp. 1070-1082.

[38] Millar, J.B., Vonwiller, J.P., Harrington, J.M., Dermody, P.J. 1994 "The Australian National Database of Spoken Language", Proc. of ICASSP'94, Vol. 1, pp. 97-100.

[39] Nadas, A., Nahamoo, D., Picheny, M.A. 1988 "On a Model-Robust Training Method for Speech Recognition", IEEE Trans. on ASSP, Vol. 36, No. 9, pp. 1432-1436.

[40] Nolan, F. 1983 "The Phonetic Bases of Speaker Recognition", Cambridge University Press: Cambridge.

[41] O'Shaughnessy, D. 1986 "Speaker Recognition", IEEE ASSP Magazine, Oct., pp. 4-17.

[42] O'Shaughnessy, D. 1987 "Speech Communication Human and Machine", Addison-Wesley Publishing Company.

[43] Peterson, G.E. and Barney, H.L. 1952 "Control Methods in a Study of the Vowels", J. Acoustic Soc. Am., Vol. 24, pp. 175-184.

[44] Poritz, A.B. 1988 "Hidden Markov Models: a Guided Tour", Proc. of ICASSP-88, Vol. 1, pp. 7-13.

[45] Rabiner, L.R. & Schafer, R.W. 1978 "Digital Processing of Speech Signals", Prentice-Hall, Inc.: Englewood Cliffs, New Jersey.

[46] Rabiner, L.R., Juang, B.H., Levinson, S.E., Sondhi, M.M. 1985 "Recognition of Isolated Digits Using Hidden Markov Models with Continuous Mixture Densities", AT&T Technical Journal, Vol. 64, No. 6, pp. 1211-1234.

[47] Rabiner, L.R., Wilpon, J.G., Juang, B.H. 1986 "A Segmental K-means Training Procedure for Connected Word Recognition", AT&T Tech. J., Vol. 65, No. 3, pp. 21-40.

[48] Rabiner, L.R. 1989 "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proc. of IEEE, Vol. 77, No. 2, pp. 257-286.

[49] Reynolds, D.A. 1992 "A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification", Ph.D. thesis, Georgia Inst. of Technology.

[50] Reynolds, D.A. 1995 "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Trans. on Speech and Audio Processing, Vol. 3, No. 1, pp. 72-83.

[51] Rosenberg, A.E. 1987 "Evaluation of a Vector Quatization Talker Recognition System in Text Independent and Text Dependent Modes", Computer Speech and Language, Vol. 2, pp. 143-157.

[52] Rosenberg, A. E. 1991a, "Connected Word Talker Verification Using Whole Word Hidden Markov Models", Proc. of ICASSP-91, pp. 381-384.

[53] Rosenberg, A.E. 1991b "Recent Research in Automatic Speaker Recognition", in Advances in Speech Signal Processing, ed. by Furui, S. and Sondhi, M.M..

[54] Russell, M.J., Moore, R.K. 1985 "Explicit Modeling of State Occupancy in Hidden Morkov Models for Automatic Speech Recognition", Proc. of ICASSP-85, pp. 5-8.

[55] Sambur, M.R. 1975 "Selection of Acoustic Features for Speaker Identification" IEEE Trans. on ASSP, Vol. 23, pp. 176-182.

[56] Savic, M., Gupta, S.K. 1990 "Variable Parameter Speaker Verification System Based on Hidden Markov Modeling", Proc. of ICASSP'90, Vol. 1, pp. 281-284.

[57] Soong, F.K., Rosenberg, A.E., Rabiner, L.R., Juang, B.H. 1985 "A Vector Quantization Approach to Speaker Recognition", Proc. of ICASSP, pp. 387-390.

[58] Soong, F.K. and Rosenberg, A.E. 1988 "On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition", IEEE Trans. on ASSP, Vol. 36, No. 6, pp. 871-879.

[59] Stevens, K.N. and House, A.S., 1955 "Development of a Quantitative Description of Vowel Articulation", J. Acoustic Soc. Am., Vol. 27, No. 3, pp. 484-493.

[60] Stevens, K.N. and House, A.S., 1963 "Perturbation of Vowel Articulations by Consonantal Context: an Acoustical Study", J. of Speech and Hearing Research, Vol. 6, No. 2, pp. 111-128.

[61] Stevens, K. 1971 "Sources of Inter- and Intra- speaker Variability in the Acoustic Properties of Speech Sound", Proc. of 7th International Congress of Phonetic Sciences, pp. 206-232.

[62] Stevens, K. 1972 "The Quantal Nature of Speech: Evidence from Articulatory-Acoustic Data", in Human Communication: A Unified View, E. E. David & P. B. Denes (eds), McGraw-Hill: New York, pp. 51-66.

[63] Stevens, K. 1989 "On the Quantal Nature of Speech", Journal of Phonetics, Vol. 17, No. 1/2, pp. 3-45.

[64] Tishby, N.Z. 1991 "On the Application of Mixture AR Hidden Markov Models to Text Independent Speaker Recognition", IEEE Trans. on Signal Processing, Vol. 39, No. 3, pp. 563-570.

[65] Titterington, D.M. 1985 "Statistical Analysis of Finite Mixture Distribution", published by John Wiley & Sons Ltd.

[66] Tohkura, Y. 1987 "A Weighted Cepstral Distance Measure for Speech Recognition", IEEE Trans. on ASSP, Vol. 35, No. 10, pp. 1414-1422.

[67] Tosi, O., Oyer, H., Lashbrook, W., Pedrey, C., Nicol, J., and Nash, E. 1972 "Experiment on Voice Identification", J. Acoustic Soc. Am., Vol. 51, No. 6, pp. 2030-2043.

[68] Tosi, O. 1979 "Voice Identification: Theory and Legal Applications", University Park Press: Baltimore.

[69] Wakita, H. & Fant, G. 1978 "Towards a Better Vocal Tract Model", STL-QPSR 1, Speech Transmission Laboratories, K. T. H., Stockholm, pp. 9-29.

[70] Wolf, J. J. 1972 "Efficient Acoustic Parameters for Speaker Recognition", J. Acoust. Soc. Amer., Vol. 51, pp. 2044-2055.

[71] Wood, S. 1979 "A Radiographic Analysis of Constriction Locations for Vowels", Journal of Phonetics, Vol. 7, pp. 25 - 43.